DOCUMENT RESUME

ED 187 930 CE 025 629

AUTHOR	Spirer, Janet E., Ed.
TITLE	Performance Testing: Issues Facing Vocational Education. Research and Development Series No. 190.
INSTITUTION	Ohio State Univ., Columbus. National Center for Research in Vocational Education.
SPONS AGENCY	Bureau of Occupational and Adult Education (DHEW/OE), Washington, D.C.
BUREAU NO	498NH90003
PUB DATE	[80]
CONTRACT	300-78-0032
NOTE	192p.
AVAILABLE FROM	National Center Publications, The National Center for Research in Vocational Education, The Ohio State University, 1960 Kenny Rd., Columbus, OH 43210 ($11.00)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	*Educational Philosophy; *Legal Responsibility; *Performance Tests; *Program Implementation; *Test Construction; Testing; Testing Programs; *Vocational Education

ABSTRACT

Addressing issues facing vocational education on the topic of performance testing, this handbook consists of a collection of seventeen commissioned papers and reactions to the papers. Two papers are presented on each of the following types of issues that must be considered before a performance test can be constructed: philosophical, technical, legal, and implementation issues. Authors were selected to form a multidisciplinary group to address each issue, and a reaction to the two papers presented on each issue is included. The two papers on philosophical issues are authored by Henry Borow and Jack C. Willers; reactions are given by John F. Thompson. The two papers on technical issues are authored by Evelyn Perloff and Raymond Klein; reactions are given by Samuel A. Livingston. The two papers on legal issues are authored by Paul L. Tractenberg and Diana Pullin; reactions are given by William G. Buss. The two papers on implementation issues are authored by H. Brinton Milward and Curtis R. Finch; reactions are given by Janet E. Spirer. Finally, two papers are included that discuss the implications of the contents of all the papers for vocational education. These papers are authored by Robert E. Spillman, Charles D. Wade and Nellie Carr Thorogood; reactions are given by Marvin R. Rasmussen. (BM)

Research and Development Series No. 190

PERFORMANCE TESTING: ISSUES FACING VOCATIONAL EDUCATION

compiled and edited by

Janet E. Spirer

The National Center for Research in Vocational Education
The Ohio State University
1960 Kenny Road
Columbus, Ohio 43210

THE NATIONAL CENTER MISSION STATEMENT

The National Center for Research in Vocational Education's mission is to increase the ability of diverse agencies, institutions, and organizations to solve educational problems relating to individual career planning, preparation, and progression. The National Center fulfills its mission by:

- Generating knowledge through research
- Developing educational programs and products
- Evaluating individual program needs and outcomes
- Installing educational programs and products
- Operating information systems and services
- Conducting leadership development and training programs

FUNDING INFORMATION

Project Title:	The National Center for Research in Vocational Education; Evaluation Handbook: Performance Testing: Issues Facing Vocational Education
Contract Number:	OEC-300-78-0032
Project Number:	498 NH 90003
Educational Act Under Which the Funds Were Administered:	Education Amendments of 1976, P.L. 94-482
Source of Contract:	Department of Health, Education, and Welfare United States Office of Education Bureau of Occupational and Adult Education, Washington, DC
Project Officer:	Paul Manchak
Contractor:	The National Center for Research in Vocational Education The Ohio State University Columbus, Ohio 43210
Executive Director:	Robert E. Taylor
Disclaimer:	The material for this publication was prepared pursuant to a contract with the Bureau of Occupational and Adult Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under government sponsorship are encouraged to express freely their judgment in professional and technical matters. Points of view or opinions do not, therefore, necessarily represent official U.S. Office of Education position or policy.
Discrimination:	Title VI of the Civil Rights Act of 1964 states: "No person in the United States shall, on the grounds of race, color, or national origin, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any program or activity receiving federal financial assistance." Title IX of the Education Amendments of 1972 states: "No person in the United states shall, on the basis of sex, be excluded from participation in, be denied the benefits of, or be subjected to discrimination under any education program or activity receiving federal financial assistance." Therefore, the National Center for Research in Vocational Education, like every program or activity receiving financial assistance from the U.S. Department of Health, Education, and Welfare, must operate in compliance with these laws.

TABLE OF CONTENTS

FOREWORD

Performance testing to measure student achievement is one evaluation method being advocated by a number of groups. However, the trends toward accountability of all public programs and the advent of such movements as minimum competency testing has raised concerns with which vocational education must deal if it is to expand its use of performance testing.

Performance Testing: Issues Facing Vocational Education addresses some of these concerns. Using a multidisciplinary approach, seventeen persons were selected to provide their views on one of four issue areas—philosophical; technical, legal, and implementation—and the implications of these issues for vocational education. The multidisciplinary approach resulted in providing a mix of thoughts which are designed to leave the reader with some new ideas and other ways to look at some old ideas.

The National Center expresses its appreciation to the seventeen contributors to the handbook: Henry Borow, University of Minnesota; William G. Buss, University of Iowa; Curtis R. Finch, Virginia Polytechnic Institute and State University; Raymond S. Klein, National Occupational Competency Testing Institute; Samuel A. Livingston, Educational Testing Service; H. Brinton Milward, University of Kentucky; Evelyn Perloff, University of Pittsburgh; Diana C. Pullin, Center for Law and Education, Inc.; Marvin R. Rasmussen, Portland Public Schools; Stephen J. Slater, Oregon Department of Education; Robert E. Spillman, Kentucky Bureau of Vocational Education; John F. Thompson, University of Wisconsin-Madison; Nellie Carr Thorogood, San Antonio College; Paul L. Tractenberg, Rutgers University; Charles D. Wade, Kentucky Bureau of Vocational Education; and Jack C. Willers, George Peabody College.

J. Stanley Ahmann, Iowa State University, Kenneth Eaddy, Vocational-Technical Education Consortium of States, and William Osborn, Human Resources Research Organization, provided useful suggestions on an earlier draft of the manuscript.

The National Center is particularly indebted to Janet E. Spirer who edited this handbook and directed the project with assistance from Nancy F. Stephens, program assistant and Ron Schilling, graduate research associate. Recognition is also due to N. L. McCaslin, associate director for evaluation and policy and Floyd L. McKinney, program director, for their assistance throughout the project. In addition, appreciation is extended to Nancy Powell and Carolyn Hamilton who typed and edited the manuscript, respectively.

On behalf of the National Center, I want to express appreciation to the Bureau of Occupational and Adult Education, U.S. Office of Education, for sponsoring this evaluation handbook.

Robert E. Taylor
Executive Director
The National Center for Research
in Vocational Education

PREFACE

A Bit of History[1]

It was almost a century ago that the infant science of psychology began to put into serious practice Alexander Pope's dictum, "The proper study of mankind of man." Wilhelm Wundt established his psychological laboratory in Leipzig, Germany, in 1878. James McKean Cattell, a young American who studied with Wundt, was convinced that the inconsistencies in the laboratory's psychological findings, which Wundt himself insisted were mainly errors of measurement, were in reality indications of important variations in human mental makeup. Pursuing his studies of human responses to simple mental tasks, first at the University of Pennsylvania around the year 1900 and, a few years later, at Columbia University, Cattell essentially launched the objective testing movement in America and is generally recognized, along with an older contemporary, Sir Francis Galton, as a founder of the subscience of the psychology of individual differences.

Early application of measurement rules to the objective and systematic observation of student achievement appeared in the work of J.M. Rice in 1897. Rice constructed a spelling test and sampled the spelling abilities of pupils in twenty-one cities. The popularity of objective tests of educational achievement to measure students' subject-matter mastery grew rapidly, and nationally standardized testing programs were subsequently adopted, not without controversy and abuses. For many decades, achievement testing took the form of paper-and-pencil tests of cognitive objectives (primarily information) of classroom instruction. Performance tests of training outcomes, as we know them today, occupied a relatively obscure place in the early history of educational testing.

A parallel development in the testing movement within psychology did, however, produce technical advances that expanded the range of testing practices in the schools. The individual mental testing methods pioneered by Alfred Binet in France proved impractical for the large-scale testing of army recruits in World War I. A five-man committee, headed by Robert M. Yerkes, was appointed by the American Psychological Association to develop a group test of general mental-ability. The product of this team effort was the Army Alpha, an instrument that proved to be an expedient way of screening people for training as officers and technical personnel.

The Army Beta intelligence tests were constructed for the testing of illiterate recruits, a device that foreshadowed the appearance of a wide array of nonverbal and manual tests. To assign personnel to such duties as cooking, baking, and mechanical maintenance, the army developed a series of oral trade tests, these representing in all likelihood the first mass use of performance like tests for purposes of certifying occupational fitness. Questions from an oral trade test for the position of machinist-die sinker illustrate the knowledge approach used: "What will happen to the dies if they are overheated and cooled too quickly?" "What is the usual finish allowance on a drop forging?" "What machine is used for cutting a straight groove between two deep holes?"

In the 1920s and 1930s, numerous tests of psychomotor abilities and nonverbal problem solving emerged, such as the Minnesota Mechanical Ability Tests. The technology of nonverbal, skill testing received further significant impetus from the efforts of the World War II army aviation psychology testing research program that produced the S.A.M. Complex Co-ordinator for the selection of military pilots. Although the evidence is not conclusive, it seems probable that the prominence of such tests was later instrumental in shifting the testing emphasis within vocational education away from the exclusive use of paper-and-pencil information measures and toward "hands-on" performance-type measures. We can be more confident about the significant impact of military personnel research during the 1950s and 1960s. The meticulous and sophisticated studies to develop and assess new performance testing procedures for technical training programs had direct relevance to the improvement of measures of student competence in vocational education.

Our View of Performance Testing

The literature is replete with definitions of performance tests and performance testing, such as:

- An applied performance test . . . measures performance on tasks requiring the application of learning in an actual or simulated setting. Either the test stimulus, the desired response, or both are intended to lend a high degree of realism to the test situation. The identifying difference between applied performance and other types of tests is the degree to which testing procedures approximate the reality of the situation in which the actual task would be performed.[2]

- A performance test is a template—a template modeled from a job task and used to gauge the similarity of a trained behavior to the demands of that job task.[3]

- In vocational and technical education the term performance test expressly denotes a measure of competency (skill level) in some specified field of occupational training . . . The performance tests may measure the test subject's handling of the work process or the quality of the work product or both.[4]

- A test of the class has designated as performance and product evaluation is one in which some criterion situation is simulated to a much greater degree than is represented by the usual paper-and-pencil test.[5]

This handbook will not offer another definition of performance testing. Rather, the authors of the papers in this handbook identified three attributes that they feel undergird performance testing in vocational education. First, *performance testing procedures attempt to approximate an actual situation drawn from a specific occupational context.* Second, *performance testing can cover some or all of the actual work situations through cognitive, affective and psychomotor domains from a process and/or product perspective.* Third, *performance testing results in a variety of outcomes, such as student certification, program evaluation, instructional planning, and information for constituencies.* Thus, the authors perceived performance testing in vocational education as an evaluative tool with a variety of possible outcome measures. It differs from other types of testing in that a performance test assesses a portion of all of an actual work setting by attempting to approximate the actual work setting.

Need for the Handbook

The need for this handbook arises from a variety of sources. For example, the stress on accountability in publicly funded programs is reflected in the federal rules and regulations whereby state boards are required to measure student achievement by standard occupational proficiency measures, criterion-referenced tests, and other examinations of students' skills, knowledge, attitudes, and readiness for entering employment successfully. Simultaneously, educators are attempting to respond to the perceived ineffectiveness of evaluation efforts to date by more closely matching the information needs of decision-makers to the evaluation questions asked and methods used to gather and interpret the information. In response to these trends and others, this handbook was designed to help teacher educators and state and local education agency personnel respond to their evaluation responsibilities.

The Approach

This handbook consists of a collection of commissioned papers and reactions to the papers that focus on four types of issues that must be considered before a performance test can be constructed. The issues include: Philosophical Issues, Technical Issues, Legal Issues, and Implementation Issues. And, two papers are included that discuss the implications of the contents of all of the papers for vocational education.

In designing this handbook, we have compiled a multidisciplinary group of authors to address each issue area. Because the issue areas themselves are broad, our space is limited, and the authors are drawn from diverse disciplines, you may find that the authors did not address all relevant aspects of the issue area. To partially compensate, we have included a Comments section for each issue area that consists of a reaction to the two papers. However, we feel that as a collection, the handbook will provide you with a foundation on issues related to performance testing, and testing in general, that must be considered before a performance test is constructed and implemented. We believe that the multidisciplinary approach will open new insights for you as you read about each issue area from these different perspectives. The mix of authors should leave you with some new ideas and other ways to look at some old ideas.

Notes

[1]Borow, H. "Philosophical, Practical, and Technical Issues Pertaining to Performance Testing in Vocational Education," unpublished manuscript, (Columbus, Ohio: The National Center for Research in Vocational Education, 1979), pp 1-3.

[2]Sanders, J.R. & Sachse, T.P. "Applied Performance Testing in the Classroom," *Journal of Research and Development in Education*, 1977, 10(3), 92-104

[3]Osborn, W.C. *Developing Performance Tests for Training Evaluation* (Professional Paper 3-73). (Alexandria, VA: Human Resources Research Foundation, 1973).

[4]Borow, H. "Philosophical, Practical, and Technical," p. 4.

[5]Fitzpatrick, R. and Morrison, E.J. "Performance and Product Evaluation." In R.L. Thorndike (Ed.). *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education, 1971. p. 238.

CHAPTER ONE

INTRODUCTION

Before discussing the four issues facing performance testing—philosophical, technical, legal and implementation—a brief discussion of performance testing itself is the logical place to begin. Stephen Slater provides an overview of performance testing in Chapter One. He begins with a discussion of performance testing focusing on the "range of test stimulus characteristics, response characteristics, and surrounding conditions, illustrating the distinctions between performance tests and other kinds of tests." A typology of performance tests and a discussion of advantages and disadvantages follow. The remainder of the paper is focused on classifying testing purpose, technical considerations, and cost considerations with performance testing.

Introduction to Performance Testing

Stephen J. Slater
Oregon Department of Education
Salem, Oregon

The person who is considering using performance tests in vocational education is faced with a staggering array of questions for which there are no easy answers. What constitutes a valid measure of occupational competence? What types of tests are most useful for guiding instruction? For evaluating program outcomes? What testing procedures result in the most reliable data? How does one begin to develop an instrument when none exists? What criteria should be used in evaluating tests developed by others? This is just a partial list.

Educators in other fields are also wrestling with these questions. The net effect is that educational measurement is currently experiencing a period of change and reconceptualization perhaps unprecedented since the days of Alfred Binet and James Cattell. While once the standardized, norm-referenced objective achievement test modeled after Binet's, Cattell's and others' instruments were held in high regard, that unquestioned acceptance is eroding. Today we are witnessing a broadening of testing options that is raising issues at a faster rate than they are being resolved. A common thread running through these options is the complexity of human competency—and the inadequacy of the ubiquitous multiple-choice examination as a measure of competence.

This chapter examines one facet of the testing options available to educators—performance testing. Throughout this examination, we seek to provide the reader with a few answers to the questions posed at the outset.

What is Performance Testing?

Defining the meaning of the term "performance test" is not as straightforward as it might seem. As with any term in our language, its meaning has shifted over the course of time and also in the way it has been used in specific contexts. For example, in the context of testing general mental abilities, the label traditionally refers to tasks requiring a nonverbal response such as arranging pictures in a logical sequence. In the armed forces, performance tests have been synonymous with measures of psychomotor skills such as speed in putting on a gas mask or disassembling a rifle. The purpose of this section is to propose a definition that conveys the current meaning of performance testing in the field of educational measurement.

Cronbach defines "test" as ". . . a systematic procedure for observing a person's behavior and describing it with the aid of a numerical scale or category system."[1] The big variable in this definition is how the term "behavior" is operationalized; doing so prescribes the characteristics of the stimulus eliciting the behavior, the type of response called for, and the conditions under which the behavior is displayed. Operationalizing behavior in these three respects is a heuristic technique for distinguishing between performance tests and other kinds of tests.

In the following pages, performance tests refer to tests in which the test stimulus, the desired response, and the surrounding conditions approximate the reality of an actual situation drawn from a specific occupational or role-based context. As implied in the word "approximate," there are many alternative approaches to performance testing, ranging along the scale of their relative realism with respect to real life situations. The following paragraphs discuss the range of test stimulus characteristics, response characteristics, and surrounding conditions, illustrating the distinctions between performance tests and other kinds of tests.

Stimulus Characteristics of Tests. Any test contains a set of instructions, a prompt, a demand, or an event that initiates the examinee's behavior. In essence, the stimulus presents the examinee with a task that can be simple or complex, structured or unstructured, ambiguous or unambiguous. The stimulus can also vary in its fidelity or resemblance to naturally occurring, real life stimuli.

For example, a student in an emergency medical technician training course might be seated at a telephone and receives a simulated (role-played) call from a parent whose child has just swallowed an unknown quantity of medication. The parent is nearly hysterical, so the student has some difficulty eliciting the necessary information (e.g., the name of the medicine and the address) and explaining what should be done before the rescue team arrives. This type of stimulus might be contrasted to a set of multiple choice questions (that might have been administered earlier in the course) measuring knowledge of appropriate emergency treatments for different types of poisoning, procedural steps in eliciting information, and how to relay information to rescue personnel.

The simulated telephone call draws upon the student's knowledge in these areas, but it also tests something more: the ability to respond appropriately in an unstructured, stressful situation. The stimulus, in its relatively high fidelity to similar real life occurrences, evokes behavior that otherwise not may be observed.

Response Characteristics of Tests. A major distinction to be made among response categories is McClelland's respondent/operant dichotomy.[2] Respondent behaviors are structured in advance by requiring the examinee to choose from among a limited set of clearly defined response alternatives. Operant responses, on the other hand, are characteristic of behavior in real life situations in which there are no artificial, preconceived constraints limiting the behavior that might be observed. Operant behavior, therefore, reflects the response capabilities of the individual as elicited by the particular test stimulus. Respondent behavior is "filtered" by the allowable responses inherent in the test. As pointed out by Paul Pottinger, competence measures allowing only respondent-type behavior are analogous to measuring how fast people can drink while requiring them to use a straw. "In this example, the paper and pencil test and the straw are equivalent in that they both limit the phenomenon being measured in a reliable way."[3]

An example of a test permitting operant behavior is to give a student pilot a chance to land an airplane. The ensuing psychomotor responses and use of judgment are unconstrained by any inherent test characteristics. A test measuring respondent behavior might pose a series of conventional multiple-choice questions concerning appropriate corrections for side winds, when to apply power in a landing, and right-of-way rules.

Choosing to test for operant behavior leads to a further decision—whether to observe the behavioral *process* or the *product* resulting from an examinee's behavior. For example, one might assess proficiency in troubleshooting a malfunctioning automobile engine by observing the sequence of steps the examinee takes in isolating the problem. The examiner might be interested in the efficiency with which the task is carried out, whether safety precautions are followed,

whether diagnostic information is interpreted correctly, and so on. On the other hand, the examiner may only be concerned with the outcome or product of the task: whether the malfunctioning part is in fact identified. The choice between process and product assessment is influenced by a number of considerations such as testing purpose, nature of the task, and relative costs of each approach. These considerations will be explored further in a later section.

Surrounding Conditions. Closely related to the stimulus characteristics discussed above are environmental conditions under which a task is performed. McGuire has pointed out that behavioral assessment in a naturalistic setting is often affected by the "accidents of nature and the flow of real problems available at the particular place and the specific moment in time when an assessment is to be made."[4] This point brings us back to Cronbach's definition of a test as "a systematic procedure . . ." The "noise" always present in reality can lead to "unsystematic procedures if care is not exercised in either of two respects: (1) standardize the surrounding conditions so it is possible to avoid confounding stimulus characteristics with irrelevant environmental conditions, or (2) systematically sample relevant surrounding conditions, building them into the test itself as variations in the test stimulus. The former condition is typically easier to satisfy than the latter, but often it is impossible to do either. In such a case, one must make the assumption that uncontrolled situational variables do not bias the description of behavior.

An example of how environmental characteristics can influence behavioral assessment is the classic case of evaluating student teachers' performance. The college supervisor making the rounds to observe several preservice teachers may notice a distinct pattern in how closely different ones adhere to their lesson plans. On one extreme, several seem never quite to make it through the rudimentary concepts they want to get across. Another group breezes through its planned activities, and the students are busily engaged in self-initiated projects. How does the student teaching supervisor take into account the fact that the former group is assigned to inner-city schools while the latter is located in suburban schools surrounding the university?

The three dimensions discussed above illustrate the ways in which performance tests differ from traditional paper-and-pencil achievement measures. They also provide a framework for describing variations in performance testing approaches and analyzing relative advantages and disadvantages of alternative approaches. The next section proposes a typology of performance testing approaches based on their relative fidelity to real life situations.

A Typology of Performance Tests

Conceptual distinctions can be made among three primary types of performance testing approaches: direct assessment, work sample methods, and simulation techniques. Each encompasses a variety of measurement options and each has its own particular advantages and disadvantages, affecting the choice of when to use a given approach.

Direct Assessment. The highest fidelity that can be achieved in assessing behavior required for success in a real life setting is through direct observation of behavior (or its outcomes) in that setting. Stimulus and response characteristics of the test and the surrounding conditions are assumed to be equivalent to those present in naturally occurring situations. Behavior exhibited in an actual work setting can be described in a variety of ways: the observer may use a rating scale to record judgments of the individual's effectiveness in a number of dimensions, the observer may record the presence or absence of predetermined behaviors on a checklist, or the observer

may count the frequency with which the individual exhibits a particular type of behavior in a given time interval. Direct assessment of products or outcomes also can rely on rating scales or checklists in which the results of the individual's performance is judged.

Direct assessment can vary in its obtrusiveness; that is, the individual may or may not be aware that his or her performance is being (or will be) evaluated. This constitutes an important advantage for the technique, relative to the other performance testing approaches discussed in this chapter. To the extent that the behavior is exhibited in an ongoing, nonartificial environment, unobtrusive observations can be made of how individuals do perform as opposed to how they can perform. Often it is not ethical or feasible to employ unobtrusive measures, but direct assessment methods do afford the opportunity by virtue of the fact that environmental conditions are not manipulated for the sake of performance testing.

An example of direct assessment is the case mentioned above where student teachers are observed in their actual classrooms. Another example is the behind-the-wheel driving test administered to drivers' license applicants. A third example is the evaluation of interns in clinical settings. Product evaluation as a direct assessment method is exemplified by judging a finished piece of work done by an apprentice plumber such as determining the watertightness of pipes joined together. All examples are characterized by nonmanipulation of the stimulus and environmental characteristics surrounding the situation in which the performance is observed.[5]

Work Sample Methods. Evaluation of work samples is distinguished from direct assessment techniques primarily on the basis of where the performance is observed. Whereas direct assessment takes place in the setting where the behavior is normally displayed, work samples can be obtained in a more controlled setting.

A second distinguishing feature is the examiner's ability to prespecify the task. Under a direct assessment approach, tasks presenting themselves to the individual are not manipulated by the examiner; in a work sample measure, on the other hand, the intent is to standardize tasks across examinees.

A third distinction is the time frame in which the task is performed. Direct assessment methods do not impose time limits for task performance, but work samples are often standardized in terms of time allowed for task completion.

Direct assessment and work sample methods share certain common features as well. The tools, materials, and other resources the examinee works with are equivalent to those available on the job. Tasks given to the examinee are equivalent to tasks performed in real life settings. In terms of the test characteristics discussed above, work samples have high fidelity to real life tasks in the stimulus and response dimensions, but surrounding conditions tend to be somewhat artificial. Furthermore, even though the test stimulus mirrors that found in the actual workplace, it is in fact controlled and specified by the examiner, enhancing replicability of the task across examinees.

Examples of work sample techniques abound in vocational education. The Plymouth Troubleshooting Contest is a case in which a discrete set of auto mechanic skills is assessed under standardized conditions. Here the task is specified in advance, requiring contestants to pinpoint the source of trouble in a malfunctioning automobile using whatever procedures they deem appropriate. Their score is based on speed in locating the defective component; as such, this is an example of product evaluation.

A second example of a work sample test is the Seashore-Bennett Stenographic Proficiency Test administered to prospective secretaries.[6] In this test, a taped voice dictates five business letters of varying lengths at different speeds. Examinees are given thirty minutes to transcribe their shorthand notes. Again, this is an example of product evaluation in that the examiner is interested only in speed and accuracy.

A work sample performance test exemplifying process evaluation is the case in which preservice teachers are asked to prepare and teach a mini-lesson on a given topic to a small group of students. The performance is videotaped and later evaluated by the master teacher, the student teacher, and perhaps the students' peers. Typically, a detailed coding form is used to quantify the types of behavior exhibited, such as using different questioning strategies, giving students positive reinforcement, following up on students' responses, maintaining eye contact, and so on. In this type of microteaching work sample, the intent is to evaluate various components of overall performance for the purpose of helping the student teacher improve in his or her areas of weakness.

Simulation Techniques. As the term is currently used in educational measurement, simulation refers to the process of abstracting some aspect of reality and concretely representing it in the form of a specific task that examinees are expected to perform.[7] Simulation accounts for an enormous spectrum of performance testing approaches, varying in their degree of "abstraction" from real life situations. At one extreme, simulation overlaps with work sample methods in tasks that recreate problems and events occurring in an actual work setting. At the other extreme, simulation techniques can sacrifice some fidelity in both stimulus and response dimensions for the purpose of gaining more control over the testing situation or avoiding the costlier aspects of duplicating reality. The range of performance testing approaches labeled as simulations includes paper-and-pencil problem solving exercises, dyadic or small group role-playing techniques, man/machine interactions, computerized games, and audiovisual representations of real life stimuli to which examinees react.

In terms of fidelity to actual situations, simulation techniques cover the considerable middle ground between objective paper-and-pencil examinations and work samples or direct assessment. Unlike the latter two types of performance testing approaches that maintain high fidelity in the stimulus and response characteristics of tasks, performance tests labeled as simulations imitate but do not duplicate reality in these two dimensions. Of course, the conditions surrounding the simulated task are typically unlike those characteristics of real life situations.

Use of simulation as a formalized technique for performance evalaution is relatively recent. In contrast, use of simulation in training can be traced to the sand table military war games of the nineteenth century, if not earlier.[8] Not until World War II were simulation and gaming techniques systematically developed for assessment purposes.[9] In the years just prior to World War II, the German Army developed standardized situational tests of team performance to select and train military personnel. British and American explorations in the use of simulation for assessment were soon to follow.

In 1943, a procedure for selecting espionage agents to serve in the Office of Strategic Services (OSS) took form.[10] The central feature of the three-day OSS assessment program was the use of situational tests designed to elicit behavior predictive of performance in actual settings. Recruits were observed in several individual and group-based exercises and then rated

on such dimensions as leadership, practical intelligence, motivation, social relations, and emotional stability.[11] For example, one task required a group of six candidates to transport a heavy rock, a log, and themselves across an eight foot stream, using only a few boards (less than eight feet long), three lengths of rope, a pulley, a barrel, and whatever trees were available Another exercise, the "Manchuria Test," provided background facts to a candidate who was then to prepare propaganda designed to lower the morale of Japanese railway workers.[12]

Following the war, the use of simulation techniques to predict future performance found applications in industrial settings as well The first use of "assesssment center" techniques, as they came to be known in business contexts is accredited to AT & T where, in 1957, Douglas Bray and Robert Greenleaf initiated their longitudinal study tracing the progress and development of young managers in the company.[13] The research project began with the participation of recently hired employees in a three-day series of business games, leaderless group discussions, interviews, and an in-basket exercise.[14] Participants were rated on twenty-five behavioral dimensions and predictions were made regarding their likelihood of reaching middle management. Neither the ratings nor the predictions were released to the organization for a period of eight years, at which time those participants still at AT&T were reassessed The forebearance of the researchers in withholding the career progress predictions—which were highly accurate—enhahced credibility of the technique's predictive validity.

More recently, simulation has been used as an evaluation technique in educational settings. Beginning in the early 1960s, Christine McGuire and her colleagues at the University of Illinois Medical Center developed several simulations. They have found simulation useful in assessing four critical components of competence: observational and interpretive skills, problem solving skills, interpersonal and communication skills, and technical skills.[15] The four major types of simulation procedures used in medical education are: (1) paper-and-pencil "programmed" examinations simulating an encounter between physician and patient in which examinees' abilities in clinical diagnosis and patient management are assessed, (2) audiovisual simulations that require the examinee to describe and interpret auditory or visual information (e.g., heart and lung sounds), (3) role-playing oral interviewing exercises in which the examiner elicits diagnostic information from a trained "patient," usually used to assess interpersonal skills as well as clinical information gathering, and (4) computer-managed robots that can be programmed to present the examinee with a variety of problems and respond appropriately to different physician interventions.

These brief descriptions of simulation techniques developed over the last thirty-five years do not begin to convey the richness and variety represented by this approach to performance testing. The work mentioned above covers only a few landmark accomplishments in the assessment of complex human performance. The technology of simulation is expanding rapidly in response to the need for valid and economical predictors of competence in real world settings.

Advantages and Disadvantages of Performance Testing Approaches

So far this chapter has introduced the critical dimensions on which performance tests differ from traditional academic achievement tests and that serve to discriminate among various types of performance tests. The preceding discussion has also proposed a three-part typology of performance testing approaches, illustrated with specific examples. However, the foregoing has not directly addressed factors affecting the use of performance tests.

The premise guiding this chapter is that any given measurement tool—whether a direct assessment method, work sample, simulation, or objective achievement test—is neither inherently valuable nor inherently worthless. Each is suited to a particular set of testing purposes, possesses different psychometric properties, is meant to measure different types of behavior, and requires greater or lower resources in test planning, development, administration, and scoring. The selection and use of a specific instrument is carried out with any rationality only when factors such as these are considered.

A danger to be avoided in adopting any measurement approach is to overemphasize one test evaluation standard at the expense of other relevant criteria. The following represents one reasonably comprehensive way of analyzing the utility of performance tests. The intent is not to promote one performance testing approach over others, but to point out the crucial questions that should be addressed. These considerations are discussed in three categories: (1) interaction of testing purpose with testing approach, (2) technical considerations (i.e., reliability and validity), and (3) cost considerations.

Testing Purpose. One of the more powerful factors influencing the design and choice of a measurement tool is the purpose for which data are being collected. Test use in education spans a variety of purposes. The most rudimentary way of classifying testing purposes is to ask the following questions:

- Are test scores sought for individual students or will scores be aggregated across students?
- What kinds of decisions will be influenced by the test data?

Four conceptually distinct testing purposes can be identified, representing different ways of answering these two questions.[16] These are: (1) formative program evaluation, (2) summative program evaluation, (3) instructional management and decision making, and (4) student certification. The former two are characterized by test score aggregation across individual students, while the latter two call for the collection and interpretation of individual student data. All four testing purposes affect different kinds of decisions. These are discussed briefly in the following paragraphs.

Formative program evaluation is conceived as an integral part of the process of curriculum development and improvement. Formative evaluation provides answers to questions posed by the developers of a program—answers that serve to pinpoint its strengths and weaknesses. As pointed out by Cronbach, formative evaluation is ". . . used to understand how the course produces its effects and what parameters influence its effectiveness."[17]

The goal of summative program evaluation, on the other hand, is to confront the question of a program's overall merit, relative to its competition. The results of summative evaluation are directed toward those who control the decisions about support and adoption, rather than toward the developers of the program. Whereas understanding the reasons for a program's success or failure is the goal of formative evaluation, ". . . understanding is not our only goal in evaluation. We are also interested in questions of support, encouragement, adoption, reward, refinement, etc. And these extremely important questions can be given a useful, though in some cases not a complete, answer by the mere discovery of superiority."[18]

The premise underlying testing for the purpose of instructional management and decision making is the notion that group-based instruction within a fixed curriculum using invariant teaching strategies does not enable each student to reach his or her highest level of learning.[19] Rather than serving the purpose of sorting students relative to their peers, student testing is increasingly being used to design and redesign each individual's instruction to promote mastery of the learning task. Instructional management, conceived in this way, requires the integration of testing and instruction, in which the teacher is provided with precise descriptions of each student's learning as a guide to modifying the instruction.

Testing for student certification refers to the practice of conferring institutional rewards (e.g., diplomas, documents certifying competence, advanced placement in a course sequence) on the basis of test performance. This use of test data is gaining considerable attention as a result of minimum competency testing programs enacted in several states and local school districts as well as the "early exit" examinations administered in California and Florida high schools. The rationale behind testing for student certification is that "seat time" is inadequate as a proxy for student competence and, therefore, more objective evidence of student achievement is necessary to restore meaning to the diploma.

How is the selection of a performance testing approach related to the testing purposes discussed above? First, one can argue that both formative program evaluation and instructional management require student performance data not just on achievement of terminal course objectives but also on "enabling" objectives—skills that constitute necessary but not sufficient conditions for success in achieving ultimate course goals. The intent is to identify points at which the instructional program is faltering, either across all students (in the case of formative evaluation) or with respect to an individual student's learning (in the case of instructional management).

Testing for achievement of enabling objectives implies the need for process measures as opposed to product evaluation, although there will tend to be exceptions to this rule. For example, at the end of an auto mechanics course, students might be expected to troubleshoot a specified set of mechanical defects. The enabling skills would include knowledge of basic engine principles and functions, knowledge of the interrelations among engine components, ability to interpret information about an operating engine, ability progressively to narrow down the most likely problems, and proficiency in integrating multiple types of information. A testing approach that would indicate student deficiencies in such enabling skills might include a set of work samples scored from a process evaluation perspective, supplemented with paper-and-pencil test items measuring knowledge of basic facts and principles. By comprehensively testing the enroute course objectives, the instructor can avoid wasting time reteaching skills already learned or neglecting to teach essential skills not mastered by one or more students.

Summative program evaluation and student certification, on the other hand, both call for a product evaluation approach when feasible.[20] This position is taken for the following reasons. First, the decision maker is interested in knowing whether students achieved the objectives stated as end-of-course outcomes; knowing why students failed to meet these performance standards is of lesser importance. Second, performance testing is an expensive undertaking under any circumstances; by focusing student evaluation only on terminal objectives and scoring performance from a product perspective, valuable resources are made available to do a better overall job of testing. Third, product evaluation tends to yield more reliable scores than those made on the basis of fleeting observations. Often, a task results in a durable product that can be judged by multiple evaluators or described in objective terms. For example, the ability to grind a machine part to a prescribed tolerance can be objectively (and reliably) scored, whereas the psychomotor skills leading to that product are more subjectively judged.

The latter point—reliability of scores—relates to another type of interaction between testing purpose and the choice of a measurement approach. The key question is how crucial—and irreversible—is a particular use of test data. In the case of student certification, the answer is obvious: Decisions made for this purpose affect individual students in important, relatively permanent ways. Any test data supporting these decisions must be highly valid and reliable. The remaining three testing purposes are most likely ranked on this dimension in the following order: summative program evaluation (because program continuance/termination is a major, often irreversible decision), formative program evaluation, and instructional management (ranked last because diagnostic information about a given student is typically supplemented with other types of data and the effects of inaccurate test data are relatively impermanent).

Since crucial and irreversible decisions based on test data demand evidence of high validity and reliability, how do these requirements affect the selection of a testing approach? With respect to validity, whatever testing approach is used should measure the skills it claims to measure at the level of complexity and sophistication at which they are taught—and learned. For example, to certify student competence in a computer programming course, a test should determine whether students can actually write and debug a program at a given level of complexity. Multiple-choice items or other types of respondent measures (if these constitute the entire certification exam) are not likely to reflect the intended course outcomes in their entirity. Beyond specifying a testing approach possessing face validity and content validity as a measure of terminal course objectives, it becomes an empirical question as to which type of performance test is the most valid predictor of competence as a computer programmer. Direct assessment, work samples, and simulation all would seem to hold no a priori advantages over one another in terms of predictive validity. It is largely the care with which a performance test is constructed and administered that determines its predictive validity. The reader interested in specific test development steps that are necessary in creating valid performance tests is encouraged to consult Klein's chapter in this volume.

With respect to reliability, product evaluation tends to produce more consistent scores across multiple raters than those obtained through process evaluation approaches. Standards for judging products or tangible outcomes tend to be more objective; hence, such judgments should be more reliable. On logical grounds, it is simply more difficult to specify the appropriate steps leading to a given product than it is to specify the desired characteristics of the product or outcome itself. If examinees can take a variety of routes in completing a task it is presumptuous in many cases to argue that one procedure is inherently superior to the others.[21]

In selecting between operant and respondent measures, the issue of reliabilty presents the test user with a perplexing problem. Multiple-choice tests of respectable length (e.g., twenty items) routinely yield reliability coefficients in the .80 to .90 range. Users of performance tests in which behavior is observed and rated by two judges are very pleased when the interrater reliability coefficient exceeds .60. Faced with a choice between a highly reliable objective test and a moderately reliable performance test, what is the test user to do? Go with the reliable but less-than-valid respondent measure—or opt for the converse psychometric configuration of a direct assessment or work sample technique? Both indices of test quality need to be weighed carefully when important decisions are going to be affected by test results. This author would argue for using the more valid measure and then taking all possible steps to boost reliability. However, this is an oversimplified response to a complex dilemma.

Technical Considerations. The psychometric properties of validity, reliability and objectivity that pertain to any measure of human behavior can be used as a framework for analyzing the advantages and limitations of performance tests. The present section extends the foregoing

discussion by focusing on the relative strengths of direct assessment, work sample methods and simulation techniques in these respects. Suggestions are also offered for increasing the validity and reliability of performance tests. A complete treatment of the psychometric considerations related to these performance testing approaches—as well as a review of pertinent empirical studies of reliability and validity—is beyond the scope of this chapter. The reader is referred to the chapters by Perloff and Klein in this Handbook for more thorough discussion of these technical issues.

Real life situations are difficult to control with sufficient precision to ensure of testing conditions across several examinees. Thus, as the performance testing technique approaches the reality of the criterion behavior it is intended to predict, standardization of the stimulus and surrounding conditions becomes more difficult to achieve. Fitzpatrick and Morrison, in summarizing their analysis of the reliability and validity of performance tests state:

> ". . . the more closely one tries to simulate a real criterion situation, the less reliable will be one's measurement of the performance. The dilemma of simulation is that increasing fidelity and comprehensiveness appear at least in a general way to be associated, on the one hand, with increasing validity but, on the other hand, with decreasing control and thus reliability."[22]

As these authors point out, if performance tests are based on a sample of real life performance (i.e., in direct assessment) that sample "must be taken under conditions representative to the stimuli and responses that occur in real life."[23] When these conditions vary from occasion to occasion, it is desirable to measure performance a number of times under a wide variety of conditions.

To provide a simple example, consider a behind-the-wheel driving test administered on the city streets. A test given at midday would present the examinee with a somewhat different set of stimuli—thus requiring different responses—than one conducted during rush hour. For example, "right of way" problems increase with the volume of traffic, but city congestion may preclude observing an examinee's adherence to speed limits. Observation of the same driver under varying driving conditions would tend to yield different proficiency estimates.

The problem of task standardization in direct assessment not only affects reliability (as estimated by test-retest methods) but also influences validity of the measure. That is, if the intent is to generalize from a sample of behavior taken in an actual work setting to performance in the larger domain of relevant tasks, evidence of the sample's representativeness is necessary.

Work sample and simulation techniques directly address the issue of standardizing test stimuli and surrounding conditions by controlling the extraneous factors that might influence performance. McGuire notes these advantages of simulation (in the context of assessing competence of health professionals) in the following ways:

> *Predetermination and preselection of the task.* It is obvious that simulation makes it possible to predetermine precisely the exact task which examinees are to be required to perform. Further, it is clear that, in contrast with the "noise" always present in reality, simulation makes it possible to focus on the elements of primary concern in a testing situation and to eliminate irrelevant and confusing complexities that would contaminate the assessment.
>
> *Standardization of the task.* Just as a given student can be repeatedly confronted with the same task, simulation enables an examining body to standardize the task for all examinees . . .

> In short, all examinees can be given exactly the same problem to cope with and this can be accomplished without an attack on nature.
>
> *Improved sampling of performance.* By standardizing the task and focusing on the most signficant aspects in each it is possible in a given time period to sample an individual's performance with respect to a much broader and more representative group of problems . . . which reality can rarely provide in a reasonable time frame. In carefully developed simulations, any problem, ranging from the most urgent emergency to illness spanning many years, can be collapsed into a half-hour exercise and summoned on demand.[24]

A second technical issue in performance tests that depend on observation of behavior is the problem of controlling bias in impressionistic judgments. Idiosyncratic rater biases such as leniency/stringency errors, the halo effect, and unwillingness to render extreme judgments are problems that lower reliability and cast doubt on the validity of measurement. These errors are the result of many factors that boil down to "the rater's willingness to rate honestly and conscientiously, in accordance with the instructions given to him . . . and factors that limit his ability to rate consistently and correctly, even with the best intentions."[25] For example, the rater may identify with the person being observed, resulting in an overly generous rating. This effect is particularly troublesome when the rater is that person's trainer or supervisor, who would prefer to bias the rating rather than risk reducing morale in the organization. Factors limiting raters' ability to rate accurately include lack of opportunity to observe, the covertness of the trait being rated (e.g., self-sufficiency), ambiguity of the quality to be observed (e.g., supervisory ability), lack of a uniform standard of reference on the rating scale, and specific rater biases and idiosyncrasies.[26]

These types of rating problems apply equally to direct assessment, work samples, and simulation techniques in which behavioral observation is the source of test data. In general, when objective performance standards are available (as in some types of product evaluation methods or process evaluation check lists), these problems are not pronounced. However, many performance testing approaches rely on impressionistic judgments that can lead to measurement error.

One of the most promising techniques for overcoming such types of rating error is the use of behaviorally anchored rating scales.[27] Rather than using such global scale anchors as superior, average, good, and so on, behaviorally anchored rating scales define scale points with unambiguous descriptions of observable behavior. By providing a clear definition of that trait being rated and a more objective frame of reference for judging individuals on that trait, behaviorally anchored rating scales limit raters' tendencies to subconsciously bias scores in the ways mentioned above. For further discussion of rating errors and strategies for attenuating them, the reader is referred to the chapter by Perloff in this volume.

Cost Considerations. Costs in developing, administering, and scoring performance tests constitute the greatest obstacle to their use in education. The technical problems discussed above are surmountable; expense in making good use of performance testing approaches is more difficult to avoid.

Conducting a cost analysis of various testing approaches is a tricky business. First, the test user must have clearly in mind the behavior to be measured and the appropriateness of alternative testing techniques in providing these measures. In some cases, certain testing alternatives will be ruled out at this point, regardless of cost. However, in many instances, two or more types of performance tests will be feasible and appropriate. At this point, the hypothesized

marginal gain in validity and other desired attributes must be weighed against marginal cost differences. Without empirical evidence concerning the validity and reliability of the alternatives under consideration, as well as accurate cost data, the decision will necessarily be somewhat subjective.

Usually, the purpose of testing will guide decisions regarding the amount of resources devoted to test development and administration. For example, student certification and summative program evaluation generally demand higher standards of validity and reliability, that are translated into higher costs. When serious decisions are at stake, more effort must be devoted to test development activities such as (1) validating the test content against tasks performed in real world settings, that is, conducting job analyses and matching tested skills with essential skills identified empirically; (2) carefully specifying performance criteria, instructions to students, and guidelines for examiners to control for various types of measurement error; and (3) conducting reliability and predictive validity studies based on pilot test administration. Greater test administration costs are warranted in the areas of (1) increasing the number of samples of behavior obtained for a given examinee; (2) increasing the number of raters to control certain rating errors and enhance reliability; and (3) investing greater resources in the use of full scale equipment required under direct assessment or work sample approaches.

Testing for the purposes of instructional management or—in some cases—formative program evaluation generally would allow relaxing the above standards. More informal procedures in test development and administration do not necessarily obviate the advantages of performance tests over respondent tests. One could argue that many instructional activities occurring in the classroom are variants of performance tests. Students routinely turn in projects or perform tasks that are in essence performance tests. The instructor's time spent in devising and grading these assignments is traditionally viewed as an investment in instruction, rather than an added testing burden. Granted, the more sophisticated simulation techniques and work sample methods require more effort to design, but the payoffs in student learning adequately compensate for the added expense.

An interesting aspect of the cost in performance testing is the issue of test security. Test developers who market standardized achievement tests are chafing under new laws that require the release of test items to the public. This results in a greater expense in developing, norming, and statistically equating new items for nationally administered tests. Test security is not a major issue in many types and uses of performance tests. Irrespective of whether the examinee knows the content of a work sample, he still must perform the task at a certain level of proficiency. In other words, it is hard to cheat when the task is to solder electrical connections or to type a letter. The exception to this rule occurs when knowledge of specific test content is likely to give the examinee an unfair advantage.

Summary

This chapter has sought: (1) to identify the essential dimensions on which performance tests differ from traditional academic achievement tests and in so doing propose a conceptual definition of performance testing, (2) to develop a three-part typology of performance testing approaches illustrated with specific examples, and (3) to examine issues affecting the advantages and limitations of performance tests. The unstated intent of this chapter has been to promote rationality in test use. As should be apparent, we have a great deal to learn about performance testing in vocational education and in other educational fields as well. The hope is that this chapter and those that follow will advance that understanding.

Notes

[1]Lee J. Cronbach, *Essentials of Psychological Testing*, 3d ed. (New York: Harper and Row, 1971), p. 26.

[2]David C. McClelland, "Testing for Competence Rather than for 'Intelligence'," *American Psychologist* 28 (1973): pp. 1-14.

[3]Paul S. Pottinger, "Competence Testing as a Basis for Licensing: Problems and Prospects." Paper presented at the Conference on Credentialism, University of California at Berkeley, 1977.

[4]Christine H. McGuire, "Simulation as an Evaluation Technique." Paper presented at the 1976 Annual Invitational Conference of the National Board of Medical Examiners, Philadelphia, 1976.

[5]The driving test is, to some extent, manipulated by the examiner through the directions given to the examine (e.g., "make a left turn here," "park the car in that space"). However, it is classified as direct assessment because the flow of events the driver must respond to is not artifically structured.

[6]H.G. Seashore and G.K. Bennett, "A Test of Stenography: Some Preliminary Results," *Personnel Psychology* 1 (1948): pp. 197-209.

[7]Jack L. Maatsch and Michael J. Gorden, "Assessment Through Simulation." In *Evaluating Clinical Competence in the Health Professions*, ed. by Margaret K. Morgan and David M. Irby, (Saint Louis, MO: C.V. Mosby, 1978), p. 123.

[8]A.W. Pennington, "History and Classification of War Games." In *First War Gaming Symposium Proceedings*, ed. by J. Overholt (Washington, DC: Washington Operations Research Council, 1961).

[9]Christine H. McGuire, Lawrence M. Solomon, and Phillip G. Bashook, *Construction and Use of Written Simulations* (New York: Psychological Corporation, 1976).

[10]Office of Strategic Services Assessment Staff, *Assessment of Men*. (New York: Holt, Rinehart & Winston, 1948).

[11]Donald W. MacKinnon, *How Assessment Centers Got Started in the United States: The OSS Assessment Program*. (Pittsburgh, PA: Development Dimensions International, 1974).

[12]Ibid.

[13]D.W. Bray and D.L. Grant, "The Assessment Center in the Measurement of Potential for Business Management," *Psychological Monographs* 80 (17, Whole No. 625, 1966). D.W. Bray, R.J. Campbell, and D.L. Grant, *Formative Years in Business: A Long-Term Study of Managerial Lives* (New York: Wiley & Sons, 1974).

[14]As described by Norman Frederiksen, "An in-basket test is a rather elaborate, realistic situational test intended to simulate certain aspects of the job of an administator. It consists of the letters, memoranda, records of in-coming telephone calls, and other materials that they supposedly collected in the in-basket of an administative officer. The examinee is given appropriate office materials, such as memo pads, letterheads, paper clips, and pencils. He is told that he is the incumbent of the administrative job and that he is to respond to the materials in his in-basket as though he were actually on the job, by writing letters and memoranda, preparing agenda for meetings, writing notes or reminders to himself or anything else that he deems appropriate." Quoted in R. Fitzpatrick and E.J. Morrison, "Performance and Product Evaluation," in *Educational Measurement*, 2d ed. by Robert L. Thorndike, (Washington, DC: American Council on Education, 1971), pp. 243-44.

[15]McGuire, "Simulation as an Evaluation Technique."

[16]These are not exhaustive of all testing purposes that might be identified, owing to the open-ended nature of the second question. However, they represent four commonly espoused test uses which, for the sake of the present discussion, are sufficiently comprehensive. A more complete treatment of testing purposes in education is found in *Guidelines for Evaluating Basic Skills and Life Skills Tests* (Portland, OR: Clearinghouse for Applied Performance Testing, Northwest Regional Educational Laboratory, 1979).

[17]Lee J. Cronbach, "Evaluation for Course Improvement," *Teachers College Record* 64 (1963): pp. 672-83.

[18]Michael Scriven, "The Methodology of Evaluation," in *AERA Monograph Series on Curriculum Evaluation*, ed. by Robert E. Stake. (Skokie, IL: Rand McNally & Company, 1966).

[19]Benjamin S. Bloom, J. Thomas Hastings, and George F. Madaus, *Handbook on Formative and Summative Evaluation of Student Learning* (New York: McGraw-Hill, 1971).

[20]In some cases, the process is the product; thus, this discussion centers on cases in which an identifiable outcome results from student performance. In those cases when only process measures are feasible, students' behavior can be scored holistically (i.e., an overall judgment is made).

[21]Exceptions to this rule are not difficult to find. For certain tasks, process evaluation criteria are highly defensible and judgments are based on these criteria can be highly consistent. Efficiency is one example. Taking the case of the computer programmer's certification examination, the student who prepares a flow chart before writing the actual program should probably be judged more efficient than the student who writes the same program by trial and error methods, requiring extensive debugging procedures. The end result may be the same in both cases, but the former student arrived at it more economically.

[22]Fitzpatrick and Morrison, "Performance and Product Evaluation," p. 240.

[23]Ibid.

[24]McGuire, "Simulation as an Evaluation Technique," pp. 11-12.

[25]Robert L. Thorndike and Elizabeth Hagen, *Measurement and Evaluation in Psychology and Education*, 3d ed. (New York: John Wiley & Sons, 1969). Emphasis in original.

[26]Ibid.

[27]R.C. Smith and L.M. Kendall, "Retranslation of Expectations: An Approach to the Construction of Unambiguous Anchors for Rating Scales," *Journal of Applied Psychology* 47 (1963): 149-55.

CHAPTER TWO

PHILOSOPHICAL ISSUES

Regardless of the type of testing program used in a vocational education program, there are several philosophical issues which undergrid the selection and implementation of a testing program. For example, tests may be used to measure student achievement, teacher performance, or the performance of a program area or school district. Each of these reasons for testing has a series of philosophical issues associated with it. Chapter Two discusses some of the philpsophical issues facing vocational educators who use performance testing.

Henry Borow begins the chapter with a discussion of the tacit assumptions of testing. He then turns his attention to such concerns as problems of validity, democratic ideals, national priorities, educational payoff, the mission of schools, vocational training, open admissions, and behavioral objectives, and the relationship of each to performance testing.

The second paper reviews several concerns raised by performance testing. In raising these concerns, Jack C. Willers views performance testing as bringing "to the fore the biting theoretical issues and value conflicts plaguing education and our broader society today." He cautions that while performance testing has legitimate uses within defined limitations, the danger exists that these "limitations will be exceeded when it is called upon to provide more than it has to offer." The Chapter ends with a discussion of these two papers by John F. Thompson.

Performance Testing and Social Responsibility: An Issues Analysis

Henry Borow
University of Minnesota
Minneapolis, Minnesota

Tacit Assumptions of Testing

The use of tests to classify students, appraise their learning potential, and certify them for diplomas or occupational competence is premised upon a number of beliefs about human behavior and examination scores which are rarely made explicit. The first of these is that people differ from one another in any specifiable trait and that such trait differences can be shown to distribute themselves along a calibrated continuum. The second assumption addresses the stability of measured trait differences. The notion that an examinee will fluctuate capriciously in intelligence, mathematical aptitude, space perception, or bimanual dexterity is offensive to the test user since such chameleon-like propensities make it impossible to render a trustworthy characterization of the individual's psychological strengths and weaknesses. It should be noted that this built-in assumption about trait stability extends beyond the question of the statistical reliability of the testing instrument per se, which Perloff discusses, and is a quality with which test theorists customarily imbue the test subject himself.

Thirdly, most current tests, particularly paper-and-pencil tests, are premised on the belief that, by combining subject responses to a series of discrete items in additive fashion, we may obtain a composite indicator of the internal trait which is being assessed. While the logic of such an inference has not often been questioned by test theorists and test users, applied psychologists schooled in the Gestalt psychology tradition of Kohler and Koffka have argued that the essential wholeness of a trait is missed by aggregating small fragments of behavior. Lay critics of testing, who tend to view any human trait as an entity, as *Ding an sich*, share this skepticism.

A fourth assumption speaks to the practical import of measured trait differences, that is, our ability to make a probabilistic statement about the student's performance level in some nontest setting (for example, an advanced training program or a particular occupation) on the basis of his test scores. It is not the student's standing on the test we really wish to know but, rather, what that standing can tell about how the student is likely to perform in some training or work for which he or she is being considered. Regrettably, scores on educational achievement and performance tests are commonly viewed as definitive indices of the behavior we truly wish to know. To leave this third assumption unverified is to bypass the obligation of test validation.

The Problem of Validity

The current controyersy affecting all ability testing, including performance testing, centers on the meaning and trustworthiness of test scores and the manner in which they are used in institutional decision making. The precise quantity of scholastic and vocational tests which are annually administered in the United States is not known, but it is commonly agreed that they number in the millions. Assessment of the effectiveness of educational programs and personnel decisions which significantly affect the careers and welfare of students and prospective workers is constantly made on the basis of test results. In public forums and in the courts, insistent questions are asked about the practice of denying admission to training programs or of failing to certify candidates for job eligibility on the basis of low test scores. Are tests accurate and equitable indicators of the individual competencies we wish to know about? This is the validity question, a complex issue which is variously treated in this handbook by Slater, Perloff, and Klein.

Long-standing and deeply rooted assumptions about the intrinsic merits of academic training have made systematic inquiries about the validity of achievement tests as indicators of subsequent nonschool performance appear irrelevant. If scholastic experience, including vocational and technical education, is of value in and of itself, then the validity of any achievement test can be defined as a function of the correspondence between the contents of the test and the aims and contents of the course or curriculum it is designed to reflect. The empirical question of what educational achievement test scores can accurately tell us about students' extra-scholastic or future job performance has not often been confronted. The predictive validities of CEEB and ACT scores have, of course, been frequently examined against college grades. But how many studies carefully document the quantitative relationship between scores on such tests, or on performance tests, and consequent career behavior?

Cronbach identifies four types of test validity—predictive, concurrent, content, and construct validity.[1] Perloff's chapter, which presents a somewhat similar classification scheme, proposes a technique labeled "consistency validity" as an improvement over the classical predictive validity approach. However, it is predictive validity (called "criterion validity" in Perloff's terminology) which has commanded major attention from test researchers since the earliest decades of their century. The construction and use of intelligence, scholastic aptitude, and vocational aptitude tests have typically rested upon the rationale of predictive validation.

A similar record of vigorous validation work cannot be claimed for the field of performance testing. With the exception of the military, the U.S. Army Air Force aviation psychology research program, for example, there have been few studies on the predictive validity of performance tests, particularly where subsequent job behavior has been used as the criterion. In general, performance tests in vocational education may be said to have a high degree of content validity. Their contents seem closely matched to the specific aims and subject matter of the curriculum. Furthermore, performance tests in vocational education which take the form of work samples or job simulations, especially where mechanical, electromechanical, or electronic testing devices are involved, possess an impressive amount of so-called face validity. That is, they look strikingly similar to the actual on-the-job task to be performed by the worker. Early developers of industrial personnel tests called this characteristic of tests "verisimilitude."

Performance tests which have high face validity or verisimilitude are so compellingly convincing in appearance that vocational educators, on-the-job training supervisors, and industrial recruitment officers are tempted to accept scores derived from such performance tests as tantamount to job proficiency. In fact, in so-called competency-based instructional programs,

students scores on these tests may serve as the critical arbiter of successful program completion. And yet, if criterion-referenced training and testing are strictly assumed to imply the existence of an external standard of performance against which test behavior may be compared, then the predictive validity of the majority of current performance tests remains unknown

The failure of the typical performance test to tap relevant factors in on-the-job training behavior or bona fide job behavior may limit its capacity to furnish a comprehensive and accurate index of the student's competency. Performance tests customarily appraise an array of cognitive and psychomotor skills. Yet, the affective domain is clearly part of on-the-job performance. Successful performance in the vast majority of occupations rests at least partially on worker attitudes and personal disposition, such as pride of workmanship, compliance with rules of the workplace, quality of personal relations, dependability, and integrity A summary published in the 1950's of over 300 studies of worker failures revealed that in the majority of dismissals, transfers, or nonpromotions due to unsatisfactory work records, factors of inappropriate personality and character, including attitudes and ethics, were involved.

How might we attack this validity problem? The technique of construct validity offers a promising approach. Let us suppose that a student who has completed a welding course and done well on his terminal performance test later proves unsatisfactory as a worker because he chafes under supervision and is described by the shop foreman as an uncooperative employee who does not follow instructions or adjust to changing job routines. Suppose further that a test of job adaptability has been constructed to measure such noncognitive or personality variables as cooperativeness and flexibility. Let us now hypothesize that a selected sample of trainees, all of whom have successfully completed the welding course (and passed the performance test) but who have scored low on the job adaptability test, will subsequently be low-rated on the actual job. If correctional findings (adaptability test scores vs. supervisory ratings with the welding performance test scores held constant) confirms our hypothesis, we may conclude that the adaptability test (measuring personal adjustment to the job) has construct validity, signifying that job adaptability is a contributing factor in success on a welding job. More importantly, we have produced a demonstrably more accurate indicator of student performance by combining information from the cognitive and affective testing instruments.

Performance Testing and the Democratic Ideal

Neither coincidence nor advances in the technology of psychological measurement alone can account adequately for the rapid ascendancy of educational testing. One must look beyond the schools and understand the changes in American social philosophy wrought by rapid rates of industrial expansion, urbanization, occupational diversification, and increased geographic mobility. The traditional social and familial patterns of an earlier era which stressed class distinctions, restricted occupational selection, and movement across social class lines have weakened perceptibly. Privileged occupational inheritance and the deliberate training of the youth of select families for continuity of leadership and power was gradually replaced by a way of life which favored economic growth and productivity as national aims. Thus, the ability of the individual to contribute to a burgeoning economy through demonstrated skill took on new importance in the social selection process. Beginning about 1900, formal education increasingly gained status with early job experience and then surpassed the latter as a mechanism by which youth sought to qualify for socioeconomic advancement.

Special training curricula, legislation mandating eligibility requirements for occupational entry, and competitive examinations became the modus operandi by which the young were prepared and sorted for access to the world of work.

During this period the advocates of unrestricted growth in goods and services in a free-market economy had seen promotion of the national good as the best way to insure the welfare of the individual. What was good for America was supposed to be good for Americans—all Americans. But the transition to a more dynamic industrial society—less classbound and rewarding individual productivity—did not culminate in the attainment of the ideal democracy that some had envisioned. We learned as a nation during the lusty social reform movement of the turn-of-the-century era, again during the Great Depression of the 1930's, and more recently during the widespread turbulence and unrest of the late 1960s and early 1970s, that the meritocratic system, by which those judged best qualified to productively serve the nation's growth needs are recognized and rewarded, is gravely flawed. Equality of educational and occupational opportunity for all citizens remains a yet unattained goal, and the advancement of the human condition has not always kept pace with economic progress. Ironically, the same educational system which appeared to provide a vehicle for socioeconomic improvement came to be seen by many among the disadvantaged as a barrier to personal advancement. Educational policy in general, and minimal competency testing policy in particular, are now inextricably caught up in this national dilemma. Some of the unresolved issues attendant upon this dilemma are briefly identified later in the chapter.

National Priorities and Individual Welfare

It may be instructive to view this controversy as a conflict between the goals of optimum manpower utilization, with gross national product as the primary criterion of the nation's health, and the quite different objective of maximizing human potentialities. One seeks a rapid economic growth rate, high employment, and high levels of productive and consumption. The other implies a bottom-line belief in the virtue of human uniqueness and its cultivation through liberal education. As we have seen, the conditions which favor the achievement of either of these goals are not necessarily facilitative of the other. The market for college graduates provides an illustration. By the end of the 1960s, college students were confronting shrinking opportunities to enter many higher-level occupational fields for which, a few years earlier, they had been encouraged to prepare. Inevitably, educational program admissions policies and testing and certification practices will reflect the impact of such changing employment supply-and-demand ratios. Just as surely, the question of "For whose good—for the nation or for the individual?" must again be raised with reference to the purposes of performance testing. And predictably, there will be no confident consensus and no facile solutions.

Education Payoff—Is it Worth the Investment?

Like other institutions—government, business, and the military—formal education has witnessed a lessening of public confidence and persistent calls for proof of worth. There can be little doubt that the current demand for accountability in education has given performance testing and competency-based programming an increased measure of importance and urgency. Although education continues to occupy a modestly favorable rank in the nation's scale of institutional values, public acceptance is now less an article of faith and is more clearly dependent upon a demonstrable track record. The message seems to be: good education will be supported but ineffectual educational programs will be trimmed or eliminated. Of particular concern to some critics are the claimed economic benefits of vocational education. Is the investment in tax dollars justified? Is there a market for the graduates of occupational training programs? Does the nation face the imminent prospect of structural unemployment, underemployment and job "spillover" for tomorrow's legions of graduates? In one way or

another, such questions are insistently posed or clearly implied in federal and state educational legislation, the Vocational Amendments of 1976, and in the charges given to the regional and topical research and development centers. Moreover, the rates of economic return on the sizable investments in human capital which educational systems require are now being studied by economists through cost-benefit analysis. Thomas, in making the case for applying cost-effectiveness criteria to school programs, advocates studying "educational organizations as open systems which are linked to the total economy through a set of inputs and outputs."[2]

One response of the schools to the demands for accountability has been to confront with renewed vigor the issue of quality control in occupational education. A three-pronged attack on the problem has been mounted: (1) curriculum re-examination and reform, (2) improved techniques of instruction, and (3) improved monitoring of the effectiveness of training. Performance tests can play a significant role in all three of these approaches. It is the last of these applications, however, which appears most open to public scrutiny and most likely to attract the interest of school boards, legislators, employers, and concerned citizens' groups. And it is from these same groups that hard questions are likely to come concering the purposes and trustworthiness not only of the educational system but also of the tests, including performance tests, which are used to appraise schools and students.

Performance Testing and the Mission of the Schools

The vindication of educational testing must rest ultimately upon the efficacy that measurement devices contribute to monitoring teaching and learning in the schools. All educational achievement tests, if they are at all relevant, reflect the undergirding philosophy and aims of the schools.

A visitor from another planet might deduce a great deal about the premises and a priori value network of the conventional academic track American secondary school from a detailed study of its examination contents. He/she would discover that the typical school achievement tests emphasize mastery of verbal and quantitative systems of communication (linguistic and mathematical knowledge) and comprehension of the terminology, facts, and principles of the major formal disciplines (natural sciences, social studies, and the humanities). He/she would learn, further, that society's ready acceptance of such masteries as the indicators of subsequent success and socially responsible citizenship in the adult world resides less on a solid basis of empirical evidence and more upon a leap of faith.

If our extraterrestrial visitor inquired into our theories of learning, he/she would find that the choice of subject matter in the traditional academic curriculum derives from the theory of general transfer of training. This belief holds that the diligent study of difficult subjects like Latin, physics, and mathematics disciplines and sharpens the mind in such a manner as to facilitate the later study of any other field of knowledge. Early and broad acceptance of the validity of this theory, coupled with trust in the wisdom of professional education planners to know what is best, endowed conventional achievement tests with a special mystique and apparently immunized them against serious challenge to their authoritative status.

It must be noted that some close relationships have been reported over the years between superior performance on achievement tests and success in higher education and in the professions and government service. How much of this correspondence is attributable to a genuine causal relationship and how much to selective bias in favor of high-scoring applicants (self-fulfilling prophecy) cannot be readily determined. Mounting skepticism has been voiced

about the meaning and relevance of conventional academic tests, much of this challenge coming from advocates of ethnic minorities and of poor, handicapped, or non-English speaking children. The general charge has been that standard achievement tests ignore many socially useful skills and talents applicable outside the school and, as such, raise discriminatory barriers against the socioeconomic advancement of the atypical student. Such tests, it is claimed, are too narrowly scholastic and slight some of the important pragmatic products of training which business and industry look for.

Well-designed and program-relevant tests can correct such claimed limitations. These tests reject the normative or relative score approach to test interpretation, employing instead some empirically established external standard which can define satisfactory training attainment. This is the strategy of criterion-referenced measurement and training programs which designate specific requirements for success or competency in a skill, i.e., specified levels of mastery, are said to be "competency based."

Historically, the rationale underlying vocational education programs stands in stark contrast to that of the older academic curriculum. At the turn of this century, the secondary schools were typically elitist training centers for children of the privileged class. Vocational courses and curricula were rare. Those youth destined to enter the labor force and the trades had to acquire their work skills on the job. Large numbers of them were the targets of labor exploitation. Many of the efforts of the social reform movement of that period were directed toward mitigating the plight of this segment of the population.

Despite the extension of compulsory school legislation to cover older children, significant numbers of urban teenagers left school to find needed employment. Vocational educators pushed for occupational training opportunities in the secondary schools to counter massive dropouts and qualify young students for entry into the labor force. One group which significantly advanced the vocational reform movement was the National Society for the Promotion of Industrial Education (NSPIE). It is noteworthy that the NSPIE recognized the indispensable tie between effective programs of vocational education and career guidance services, and it was this organization which was instrumental in siring the National Vocational Guidance Association, the first national society devoted exclusively to the advancement of guidance.[3]

Given this climate of practical urgency, the philosophy of vocational education, the design of its curricula, and its approach to the measurement of student achievement developed along boldly utilitarian lines.[4] The Fourth Yearbook of the American Vocational Association, which takes the philosophy of vocational education as its theme, projects a straightforward and unidimensional image.[5] There is no detailed explication of the value roots of vocational education nor of possible philosophical agreements or quarrels with the concerns of humanistic psychology—self-actualization, student-centered education, and the debilitating psychological effects of alienation. Endorsement, however, is given to the importance of developing originality and thinking ability and to the principle of individualized instruction to accommodate wide differences in student backgrounds and learning abilities. Here as elsewhere in the literature of vocational education, a plea is made to insure that "student performance criteria (be) based as realistically as possible on occupational demands."

The simple pragmatism which permeates the avowed aims of vocational education makes it particularly receptive to performance testing procedures. Yet, since educational values and goals in a pluralistic society do not form themselves into a tidy monolith, vexing problems and unanswered questions about the concept and practices of performance testing remain. These will be noted later in the chapter.

Vocational Training as an Adverse Alternative

That college preparatory programs have over the years enjoyed favored status in the public view at the expense of occupational training has produced special problems for vocational education students and staff alike. Often considered a dumping ground and salvage operation by academic purists and elitists, vocational schools have faced a particularly arduous challenge in the conservation of undervalued human resources and in equipping their students for entry into the labor market. Given these circumstances, it is not surprising that training as a tryout experience and as a form of career guidance has held a prominent place in the goal hierarchy of the vocational schools. Thus, the literature of vocational education frequently mentions the need for appropriate evaluation techniques to monitor student progress and the efficacy of instruction.[7]

Unlike the conventional academic curricula, where student grades have been employed as general indicators of readiness for occupational entry or higher-level schooling, vocational programs have been expected to furnish clearer and more direct evidence of task mastery by students. State industry-labor apprenticeship councils and other certifying bodies now specify minimum standards of acceptable work-related behavior in terms that schools cannot afford to ignore. Some authorities now call for a detailed series of tests which will provide information about the noncollege-bound student comparable to the information which the standarized achievement test battery furnishes about the college bound. Sidney Marland, the former U.S. Commissioner of Education who later proposed career education, wrote:

> A culminating examination should be created with all the strength and quality and prestige that now characterize the College Board examinations. This examination should include, in part, the appropriate academics of a liberalizing curriculum, but it should have as its principal message a measure of the quality of skilled performance in a given occupation that may be expected of the examinee.[8]

Taking a cue from the CEEB, Marland suggested that this new type of test be called the JEEP (Job Entry Examination Program).

Open Admissions and Performance Testing at Risk

Two significant contemporary trends in American education—the open admissions policy in colleges and technical schools for disadvantaged and nontraditional applicants and the adoption of program-completion certifying examinations—appear to be on a collision course. One leads to a substantial increase in the proportion of students with marginal skills for academic survival; the other sets a uniform standard of acceptable learning and may produce an increase in student failures. Many high schools have attempted to settle the problem of low-achieving students by quietly adopting a policy of automatic promotion. Criticism of this policy has been widespread and severe. Faced with growing percentages of high school graduates who enter institutions of higher learning (now over 50 percent), our colleges have three choices: (a) grade inflation; (b) watering down the curriculum; and (c) maintaining past grading standards, testing standards, and course requirements and letting dropout and failure rates run the consequences. There is at least indirect evidence that the first two alternatives are now being widely used, although it would be difficult to find those who approve. The third alternative, although more forthright, again satisfies no one, and, in addition, creates serious embarassment for the institution. Culturally disadvantaged students who entered the institution with high hopes may feel disillusioned and betrayed by false promises and expectations when they fail. Students, parents,

and governing boards may then charge that the school does not provide useful educational services of a reasonable quality. Moreover, the prospect of wholesale test-based failure rates in a period of declining enrollments inevitably invites the institution's anxious attention to the remedial needs of the marginal students. Grant has stated the case:

> "From the institutional point of view, the major impact of adopting a competence-based approach is to shift more of an institution's resources from the best to the average and below-average students. Those 'invisible' students, formerly given C's and D's for endurance and passed along, become highly visible in a competence-based format and no longer merely slip through the institution unnoticed. The competence approach forces a redistribution of faculty labor to them. A higher proportion of the faculty will spend more time teaching these students basic skills and helping them achieve specific outcomes than in traditional schools."[9]

It is clear that competence-based education and performance testing, when used to certify student mastery of required skills and understandings, may exacerbate certain already existing problems.

Performance Testing and Behavioral Objectives

One of the most compelling and attractive features of performance testing, when linked with competency-based education, is its insistence on operationalizing instructional goals and casting them in a readily observable and quantifiable form. The task of conceiving and constructing a performance test directs specific attention to the issue of training objectives. What is it in behavioral terms, i.e., directly observable responses, that the training program is attempting to accomplish? Assuming that the student has acquired the techniques which provide the *raison d'etre* of the instuctional process, what is it in specifiable terms that the student should now be able to do and to understand? While it is, of course, true that the development of any educational achievement tests may force this kind of close look at the purposes and outcomes of instruction, this advantage seems especially true when the competence-based strategy is applied to the construction of performance tests. Beyond the question, then, of how effective a performance test may be as an instrument of appraisal, the complex act of planning and constructing it has a potentially salutary effect on the process of instruction itself.

Let us see how the logic of competence-based education underlies the development of the test. Since performance tests are not isomorphic with actual job performance but are at best analogues or predictors of the latter, test researchers and technicians have had to grapple with the question of what constitutes a workable test. They must decide what features of an evaluation device make it administratively feasible and, at the same time, allow it to approximate both the training objectives and behavior on the actual job.

A prior condition to be met, however, is the specification of the logical sequence involved in the measurement of the learning itself. In brief, these steps include (a) identification of the units of behavior which are central to performance on the job for which the training is expressly designed; (b) selection of operational criteria matched to the units of job-relevant behavior identified in the initial step of this sequence; (c) determination of what is to be learned from the formal training experience itself that will optimize prospects for the development of the aforementioned behavior units; (d) specification of the learning content and goals in step c in measurable, i.e., directly observable, terms; (e) arrangement of the conditions of training and training performance such that extraneous variables, i.e., those not pertinent to the occupation

itself, are controlled or minimized; (f) assessment and scoring of those behaviors within the training setting as specified in step d; and (g) statistically relating the data derived from steps b and f. The final step in the sequence provides both a validation index of the training criterion and, less directly, a measure of the relevance or effectiveness of training. As previously noted, this ultimate operation is rarely performed because rigorously controlled follow-up occupational data may be difficult to obtain and because, further, the validity of the performance tests used to appraise training outcomes is seldom questioned.

The Behavioral Objectives Controversy

Lively disagreement exists among educators concerning the merits of behavioral objectives in performance testing. Critics argue that behavioral objectives give clarity and specificity of educational outcomes at the sacrifice of the deeper understandings involved in learning. Reducing complex instructional goals to a series of discrete, easily measured tasks or responses, they believe, may barter some of the more distinctive products of human learning like creative thinking and imagination for trivia.

In truth, behavioral objectives often appear to be excessively lean and limited in scope. Advocates of competence-based performance criteria counter by noting the nebulous nature and inaccessibility of global objectives. Frequently, too, analyzing a complex skill into its component parts may afford a more effective means of planning the teaching of that skill and of measuring the instructional product. And for complex tasks requiring mastery of a known set of identifiable principles and psychomotor operations, as in many jobs, casting the goals of learning in behavioral form may be quite advantageous. Still, it must be conceded that the behavioral approach to identifying training objectives may give disaproportionately heavy attention to those which can be most readily transformed into directly observable and conveniently recorded responses.

A specious criticism of behavioral objectives occurs when the concept is used as a synonym for behaviorism. The behavioral objectives approach uses the behavioristic principle of specifying behavior in terms of observable responses. However, as applied behavioral technology, behaviorism goes far beyond the questions of how objectives are derived and stated. It deals with the techniques for systematic behavior intervention and change through application of such principles as classical and operant conditioning. These include positive reinforcement, aversive stimulation, counterconditioning (desensitization), and even social modeling. Since none of these techniques is applicable in generating the behavioral objectives for a performance test, it is irrelevant to attack behavior objectives qua behaviorism.

Unresolved Issues

If one accepts the thesis that competence-based education holds the promise of bringing curriculum design and educational experience closer to relevant life experience, the potential value of performance testing as a means of monitoring both the quality of the instructional process and certifying student attainment of specific goals seems beyond serious dispute. To embrace this premise, however, is to simultaneously assign increased significance to performance tests and to invite some vexing questions about the limitations of testing and inappropriate testing practices. Unless the urgency of such questions is acknowledged, the performance testing movement may flounder or lose its direction and become the target of even more strident public attacks.

A number of technical measurement problems in performance testing, including those of validity, reliability, behavior sampling, and cutting scores, are addressed in this chapter and elsewhere in the volume. What remains are certain unsettled issues concerning the interpretation of performance test scores and the proper place of such tests in improving the quality of education. These concerns will be briefly noted in the form of questions.

1. *How much bearing should performance testing have on what is taught?* As previously noted, a well-designed performance test will be keyed to teaching objectives and to the masteries to be achieved. Accordingly, it should come as no surprise to find a substantial correspondence in competency testing between test contents and the subject matter of instruction. What may occur in teaching practice, however, is a subtle reversal in antecedent and consequent conditions by which the test becomes the curriculum and the school, unwittingly perhaps, begins to teach for the test. Under such circumstances, a real danger exists that performance testing may become the basis for a new meritocracy. The best of tests offer only a limited sampling of the behaviors and competencies which schools wish to transmit. Although the aims of education may be defined in terms of test content, tests are not identical with the corpus of education. To arrange instructional experience so that only the contents of performance tests are taught would be to render the educational process static and unduly confining.

2. *How much reliance should be placed on performance tests in making educational decisions?* This question inquires tactitly about the confidence we can justifiably place in tests as indicators of students' true competence. Tests can never wholly capture the *mise en scene* in which we wish to observe the student at work and in life. While well-designed performance tests may provide one of the best means of judging a student's eligibility for training or for a vocational certificate, they fail to reproduce the full range of conditions which come into play when a student is adapting to post-school experiences, including employment. Hence, it will generally be wise to combine test information with other relevant sources of information when making judgments about a student's competence. An example would be the training supervisor's systematic and standardized ratings of a student's performance in a cooperative work setting.

3. *Does the use of performance tests tend to unduly hasten occupational program decisions by students and narrow their curricular experiences?* It is common to encourage high school students in a system of competence-based vocational education to shape their course selections to the skills and understandings they must demonstrate through testing. Yet, the career plans of many of them are still unstable. The secondary school experience should be so arranged as to present a broad spectrum of exploratory activities for students. It should facilitate the process of career development rather than close it down with occupational training which is irreversible or too restrictive.

4. *Will a trend toward increased performance testing in competence-based education discourage emphasis on the liberal arts?* When tests are used to assess a narrow band of vocational abilities, the net effect is retrogressive. The need to strengthen the vocational aspects of education so that all students leave school with marketable skills is readily conceded. Still, as Willers points out in his chapter on philosophical issues, the more specialized career goals are defensible only when they are derived from and articulated within a comprehensive system of general educational goals. It follows, then, that occupationally-oriented performance testing should bear a kinship to tests of competence which are linked to the aims of broad, general education.

5. *Does the teaching and testing of standard operating procedures and a fixed body of knowledge tend to promote rote learning and discourage creative problem solving?* Note was taken previously of the behavioral strategy of stating the performance outcomes of training in crisp, directly observable terms. This approach to specifying objectives offers obvious advantages but, at the same time, tends to load tests with fragmented, static, and closed-system contents. There is need for experimentation with performance test item types which stress broad conceptual relationships, logical reasoning ability, and originality in problem solving.

6. *Are individual students sometimes the victim of unfair decisions based solely on low performance test scores?* A qualifying examination which possesses at least moderate predictive validity will classify students (for purposes of program admission or program completion) with a degree of accuracy substantially greater than chance. Furthermore, for test applicants as a group, the average discrepancy between predicted and actual performance on the criterion measure will be significantly smaller than errors resulting from guesswork or those resulting from traditional screening interviews and letters of recommendations. For many years it has been this empirically demonstrated ability of valid tests to outperform older screening methods that has justified their use in making classification decisions about students. However, unless a test has perfect validity, a condition which never occurs in reality, some students will always be misclassified by the test scores. It has been recent challenges by student candidates who have apparently been able to show that they possessed the competency denied by their low test scores which have brought the issues of test fairness and competence-based education to public attention. As the chapters by Pullin and Tractenberg show, accountability through performance testing entails a number of thorny ethical and legal considerations, and the controversy remains unresolved. But for many test designers and users who must deal realistically with the state-of-the-art limitations of measurement devices, criticisms of competency testing often appear too severe.[10] Until more accurate methods of certifying student performance can be developed, they ask, does it not make sense to use the most accurate testing procedures available, procedures which minimize classification errors?

Notes

[1]Lee J. Cronbach, *Essentials of Psychology Testing*, 2d ed. (New York: Harper & Brothers, 1960), pp. 103-23.

[2]J. Alan Thomas, Cost-Benefit Analysis and the Evaluation of Educational Systems, Proceedings of the 1968 Invitational Conference on Testing Problems (Princeton, N.J.: Educational Testing Service, 1969), p. 90.

[3]W. Richard Stephens, *Social Reform and the Origins of Vocational Guidance* (Washington: National Vocational Guidance Association, 1970).

[4]William Micheels and Ray Karnes, *Measuring Educational Achievement* (New York: McGraw-Hill Book Company, 1950).

[5]Melvin L. Barlow, ed., The Philosophy for Quality Vocational Education Programs, *4th Yearbook* (Washington: American Vocational Association, 1974).

[6]Tim Wentling and Tom Lawson, *Evaluating Occupational Education and Training Programs* (Boston: Allyn and Bacon, 1975).

[7]Richard Erickson and Tim Wentling, *Measuring Student Growth: Techniques and Procedures for Occupational Education* (Boston: Allyn and Bacon, 1976).

[8]Sidney P. Marland, A Customer Counsels the Testers, Proceedings of the 1968 Invitational Conference on Testing Problems, p. 111.

[9]Gerald Grant et. al., *On Competence: A Critical Analysis of Competence-Based Reforms in Higher Education* (San Francisco: Jossey-Bass, 1979), p. 11.

[10]Ralph W. Tyler and Richard M. Wolf, *Crucial Issues in Testing* (Berkeley: McCutchan Publishing Corporation, 1974).

Philosophical Issues in Performance Testing

Jack C. Willers
George Peabody College
for Teachers of Vanderbilt University
Nashville, Tennessee

"A workman is commendable, not for the will by which he works, but for the quality of his performance." With this perspective, the thirteenth-century philosopher Thomas Aquinas illumined a basic value of western civilization. Even if it be true that "where there is a will, there is a way," the fundamental and final criterion of a working person is the quality of his or her work performance.

Yet there has always been another perspective, not diametrically opposed to the quality of performance, but placing its higher hopes in pure theory, in intellectual contemplation as an intrinsic and ultimate value in and of itself. The history of education, of our civilization, and of our nation is the story of the conflict of this counterperspective with the values that give highest priority to the quality of performance and product. Today, the history of that conflict between thinking and doing may be seen in the issues in performance testing.

Should educators limit themselves to the basic cognitive tasks of reading, 'riting and 'rithmetic, so that formal learning in schools will be restricted to the intellectual skills necessary for academic scholarship? Or, is the primary purpose of education to instill a sense of the competitiveness of social and economic realities and, accordingly, to prepare students to perform their best in the worst situation? Or, again, should the schools place a higher priority on recognizing the inherent worth of childhood and youth, not as periods of preparation for some unknown adult future, but as time for joy and celebration, for self-expression and good feelings about one's self?

These questions are, admittedly, phrased in ways that educational theorists would never propose for their purposes and programs. This outlandish manner, however, is not to disparage the serious enterprise of thinking critically about schools and teachers and learners. Instead of belittling the difficult but necessary task of asking hard questions about education and human development, we must at times ask them in taunting, jeering terms to reveal their underlying narrow assumptions and myopic prescriptions.

No single educational philosophy or program will meet even most of the needs among individuals in a multicultural, pluralistic society characterized by competing interests and conflicting values. Still, educators easily become infatuated with fads, infuriated by failures; inflated with easy, fleeting successes; and more often than not, infected by the infallibility of our own purposes, perspectives and programs.

All educational reforms, therefore, such as performance testing or competency-based instruction, deserve the critical review, as well as the experimental testing, that gives them the opportunity to prove and to improve their own performance. Whatever the philosophic bases for performance testing, it must at least open itself to the tests of performance and subject itself to critical evaluations according to fundamental, often conflicting and certainly competing, values.

Evaluation by performance of psychomotor, job-related skills will certainly not receive the wholehearted support of classical educators. Nor is performance testing enjoying the firm support of humanistic educators who emphasize the values of play and leisure and inner self-direction. Especially critical of performance testing today are those educators who are sensitive to the apparently destructive forces in modern technology that deplete our natural resources, pollute our environment, disturb delicate ecological balances, and exploit and dehumanize skilled working people for profit and power.

Arguments for fairness in testing notwithstanding, these critics and skeptics have legitimate messages of caution. In general, these issues speak to the limitations, narrowness and inadequacies of performance testing when overstressed or used to the exclusion of other claims and interests. Though threatening to narrow self-interests, such critical messages of caution can provide clarity and breadth of purpose together with insights into other worthy means of judging human development and achievement.

From Analytical Definition to Critical Judgment

A performance test is presumed to be a measure of occupational competency or the ability to perform a job-related skill. This presumption, in turn, is based on the assumption that job skills, and even overall occupational functions, can be reduced by analysis to meaningful, manageable and measurable sequential segments. The competent performance of these work segments may then be examined and evaluated. The purpose of performance testing, accordingly, is to discern the quality of a particular individual's competency to perform a particular job-related skill or to qualify for a particular occupation.

Human beings, accordingly, are selected for additional training, jobs, and careers, and even certified for various occupations—in other words, granted the rewards for individual effort and social usefulness—on the basis of others' critical evaluation of their competence as indicated by the measurement criteria of performance tests. This analysis is entirely different from the more straightforward proposition that people are selected and rewarded on the basis of their own actual performance on the job or in the occupation. Test designs, test criteria, job descriptions and occupational analyses, test constructors and their judgments on what to measure and how to measure it, test administrators and test evaluators all stand between the individual performer and the rewards dispersed.

Furthermore, a performance test, providing an adequate basis on which to judge the degree of proficiency with which an occupational competency is performed, must also provide quantitative measurements by which the more competent craftsman may be distinguished from the less competent. Thus, not only does the analytic reduction of work sequences underlie performance testing, but also the measurement and evaluation of job-related skill competency require the quantification of qualities of both performance and product.

A well-defined objective is essential to performance testing. That objective may be the ability to perform a manipulative skill to a certain qualitative degree, or to produce a final work product

that meets certain standards of quality control. In either instance, it is not merely the performance process or the product of the work-sample that is evaluated, but rather the competency of the work which is critically assessed.

The purpose of analyzing and characterizing performance testing in the above manner is not to establish either a working or final definition of performance testing on which all may agree. The above analysis does make clear, however, that regardless of the technical definition or the characterization of performance testing used, performance testing, like all other forms of evaluation, inevitably must make assumptions about reality and human experience. It must claim some value criteria for the discernment of quality and the judgment of degrees of quality. And, performance testing must reflect some beliefs about how we learn, about how we demonstrate and apply knowledge, and about the values assigned to that knowledge.

This perspective places performance testing squarely in the philosophical domain of critical interpretation of beliefs about reality, values, and knowledge. Presuppositions and fundamental value beliefs require identification, clarification and criticism. Basic concepts of reality, intelligence, and social utility must be questioned, or at least held critically, and applied cautiously in diagnosis, evaluation and justification. The assessment of performance competencies from this perspective is a human affair, not mechanical, not prescribed or determined, but subject to the whim and prejudice or capriciousness, as well as to the reasonable disagreements of rational people.

The argument may appear strained and unnecessary to those who already acknowledge the human elements and the concomitant possibilities of error in performance testing. On what logical or utilitarian grounds is an individual justifiably subjected to performanace testing? Are there other, better reasons for not testing performance? Is such testing a subjection to external, impersonal norms that are less valuable or substantiable than others? Or is performance testing, rather, an individual opportunity to express unique human dignity, to excel, to learn about and respect one's own self?

But these and many other philosophical inquiries do not suggest themselves to those who, with a deterministic or mechanical perspective, view evaluation in general and performance testing in particular as the automatic process of perceiving degrees of quantitative variance or correspondence between two sets of clearly observable data—the external test standards and the behavioral performance. From this perspective, no values, interpretations, judgments or responsible assumptions are expressed in constructing performance tests, evaluating their outcomes, or even in the decision to administer them. To the contrary, performance testing is a value-free maneuver, a technical operation freeing both the tester and the performer not only from capricious judgments of the quality of competency but, more significantly, also from all questions of fairness and justice in allocating economic rewards and social recognition on the basis of performance. The only problems or issues related to performance testing from this latter perspective are the technical questions of test validity and reliability.

However, even the troublesome, tentative question of whether performance testing is, on the one hand, a human interaction, consisting of purposes, intentions, social goals, culturally defined criteria, and theoretical assumptions or, on the other hand, a value-free mechanical operation, is itself a question which justifies, even requires discussion regarding the philosophical issues.

Educational Goals and Performance Testing

The aims and goals of education provide a perennial pursuit for philosophic perspective. The value of life and the values worth seeking and living for are constant questions perplexing the critical mind. This has always been true but appears even more so in an era committed to science and technology, neither of which in itself purports to define our values or to solve our value conflicts. Indeed, from one debatable view, science and technology, while claiming to be value-free, have called into question our more stabilizing, traditional values, thereby creating many of our value conflicts and dislocating core values necessary for social cohesion and continuity.

In an age in which science and technology of overpowering dimensions dominate the curriculum, what is education for? What are the values and goals sought through the myriad forms of instruction, training, programming, conditioning, teaching and testing? If there is some answer to this question, it would necessarily be complex, but even then we would have only a description of the various social and personal goals people strive to achieve through learning. More crucial is the normative question: What ought to be the aims of education? From differing responses to this primary question follow the practical matters of designing curricula, applying instructional methodologies, organizing and administering learning situations, and evaluating the results.

The question of aims, like all normative questions, cannot be answered in any final sense, only in terms of philosophic perspective to which there would be equally appealing or more or less defensible counterperspectives. This is not the place to argue for this or some other goal of education. But to place the question in terms relative to performance testing, let us at least propose a theoretical framework from which to work. This approach is attributed to Thomas F. Green and can be pursued in greater detail and accuracy in his "Minimal Educational Standards: A Systematic Perspective."[1]

The aims of education may be either general or specific. Specific educational objectives indicate that which constitutes their own achievement and also designate the time when the goals are to be achieved. In this respect, performance-based training always aims at specific goals in the form of behavioral objectives, and it is the function of performance testing to indicate when and to what extent these specific goals are attained.

General educational goals, on the other hand, are vaguely expressed so that it is never possible to discern when they have been attained or the extent of their attainment. Accordingly, no form of educational measurement, perhaps least of all performance testing, can measure the achievement of general educational goals. For this reason, and because our culture places such great emphasis on measuring and counting for the purpose of efficiency and economy, it is suggested by the accountability movement, competency-based instruction, and the efforts to manage education by objectives that all seemingly useless general goals be replaced by specific objectives, the attainment of which can be measured, monitored and managed.

But the argument to eliminate general aims, in favor of the specific, rests on a misunderstanding of the function of general goals, which is to designate, not what is the good or the best, but rather what is unacceptable. As such, the general goals of education provide the grounds for defining specific educational objectives.

In short, "general goals operate effectively in the establishment of specific targets provided we recognize that their function is to provide criteria for determining what kinds of arguments

will constitute serious charges of failure. Specific educational goals are derived from general educational goals through a social process in which there is produced a definition of what constitutes not the best, but the worst that is acceptable. To suppose that specific goals for the system can or must be generated independently of general goals is to succumb to a most fundamental misunderstanding of the nature of educational goals."

The achievement of specific aims is the business of competency-based instruction, and the measurement of that achievement of specific aims is the business of performance testing. Now, the problem remains as to whether these specific aims depend upon the more general aims of education. Or have competency-based instruction and performance testing replaced general aims with specific performance targets arising from the art of the possible in instruction and the prescription of behavioral objectives? Unless specific goals depend on the general aims of education, performance testing, and its array of associated educational movements, will drive us further into an educational malaise of confusion and lost confidence.

If specific goals are not related to more general goals that express broad social values and shared ideals, narrow interests will continue to compete ruthlessly, Unsuspecting learners, striving to improve their own performance, will be caught up in the competition to exploit their improved competency. And schools and educational systems will continue to be condemned for lack of efficiency or productivity or almost any other failure, presumed or real, on grounds which are irrelevant because they do not reflect general goals of the society or the system.

The discussion seems to have generated another dilemma for performance testing: Competency in performance cannot be tested fairly unless it is an established and measurable objective of instruction. Such an objective is necessarily specific, designating the specific criteria for the evaluation of its own attainment. The behavioral objectives of competency, furthermore, emerge directly from particular job-related skills, not from broad cultural aims and values or from general social ideals and goals. And yet, as it has been argued, it is exactly these kinds of narrow, specialized performance goals that endanger the society's sense of commonalities and, consequently, the individual's relationship to that fragmented society.

Studies regarding individual alienation and dehumanization need not be recounted to strengthen the argument against specific instructional goals sought in isolation from broad social ideals and general educational aims. But the tragic picture does flash across the screen: a highly proficient person, competent in a variety of economically useful skills, who possesses little or no sense of individual or social identity, self-worth, or meaningful direction for life. Such a person skillfully fells the trees without ever seeing or appreciating the beauty of the forest. The concomitant destruction of our physical environment and the senseless waste of our natural resources, almost matches the loss of human resources.

With respect to the dilemma, some uneasy compromise between the demands of the technical and the necessities of the human may provide some small consolation. The compromise will not satisfy those who give highest priority to the inner dignity of the person rather than to creature comforts and increases in the gross national product. But, given the present power of the continuing persistence for consumption over creativity, something by way of compromise may be better than nothing at all.

This possibility of compromise lies in the hands of vocational and technical trainers who construct behavioral objectives and utilize performance tests. To these educators and evaluators fall the opportunity at least to refer the specific goals of training to broader social and educational aims.

In what respects, it might be asked, does the attainment of proficiency in some performance respond to the broader, generally accepted goals of our culture to encourage individual creativity, to foster conservation of our natural environment, to develop critical yet cooperative citizens, and to stimulate a sense of self-identity as well as a sense of belonging? How might the assumed opposition of habituated skills and creative imagination, of routine work and expressive leisure be reconciled into mutually complementary counterparts?

To ask such questions and to begin to answer them and to apply partial answers in actual practice, requires that the vocational education evaluator become an educational sociologist, historian, and philosopher, able to recognize and critically evaluate the general aims of education in order to give meaningfulness to specific objectives. Above all else, for social concord and individual human development, it is necessary that the performance instructor and evaluator judge far more than the skilled performance, and that the learner learn far more than performance skills.

Performing Slaves—The Perennial Fear

Philosophical issues converge on performance testing from across the spectrum of educational thought, even from opposite directions. From the radical end of the continuum, neo-humanistic educators, third-force psychologists, and existential philosophers rail against imposing external standards on unique individuals who are free to choose their own values and destinies. On the other hand, educational fundamentalists, the perennialists, would return our modernized, mass, corporate culture back from the vocational training of slaves to enduring universal truths and values which serve as absolute criteria for human behavior, action and performance. For these educators, the aim of schooling is "manhood, not manpower."[3]

From this latter perspective, human performance is not to be measured in terms of individual interests or needs, for all people possess a common natural power for rationality. Education must accordingly rely on the universal and the permanent, not the particular and the transitory. Nor is human performance to be measured in terms of particular marketable vocational, technical and professional skills which, apart from the power of rational judgment, mark our society's performing slaves. Corporate industrialization, technological advances, and the observation of changing facts, all served by performance training and testing, readily enslave the skilled in whom the potentiality for rational self-direction remains unrealized.

Thus, Robert Hutchins, in advocating perennialism in education, rejected outright most of the commonplace objectives of American schooling today, and especially training in vocational competencies. Since a system of education will invariably reflect major cultural forces, he argued, it would be naive to think that the schools could develop intelligent humans when all social pressures are applied to the development of uncritical, unthinking consumers and producers. Our cultural mission must, therefore, be redirected, away from national power and accelerated technological changes that take no thought of rational human progress or social consequences, toward wisdom, understanding, intelligence, and rational thought and judgment.[4] To realize our rationality and thereby reach our full human potential, a liberalizing, freeing education must be provided to all, "not to make practitioners but to help in the development of intelligent men and women."[5]

One could probably argue well that there is nothing in performance training and testing that is inherently contrary to intellectual development itself. And tests can and have been intelligently developed that do measure performance abilities and competencies. But, then, those who make such successful arguments, and those who construct such reliable and valid performance tests,

must be utilizing a degree of rationality that the skilled performer may not have had opportunity to develop. And this possibility is the crux of the issue. While it is not a matter of either rational judgment or skilled performance, it is a question of priorities. Skilled performance without critical, rational intelligence becomes, in a world of rapid technological change and built-in obsolescence, a prelude to participating in one's own victimization. Just as education reflects the dominant forces of the culture, so we teach toward the tests. And performance testing presents a danger of luring job-skills with instant yet transitory reward.

Still, it may be argued, it is better to possess any marketable performance skill than none at all or, what may be worse, an impractical, purely contemplative intellect (if there is such a thing). But this argument forces us back into an either-or dichotomy by denying all other alternatives to the extremes—either the performing slave or the intellectual who would be free if he/she knew how to do anything at all other than think his/her own thoughts. But these two alternatives are far from exhausting our human possibilities, and, besides, performance often depends on creative or critical judgment and cognitive knowledge that no strict performance test alone can measure.

The intelligent, creative and critical worker is, therefore, no threat to vocational education or performance testing. Rather she or he is the challenge.

Self and Society—the Continuing Split

Performance of skills and evaluation of performance may be viewed from the perspectives of three domains commonly used today to classify educational objectives—the cognitive, the psychomotor, and the affective. Performance testing is primarily, though not exclusively, concerned with the measurement of the achievement of psychomotor objectives and competencies. As we have seen, educational fundamentalists are concerned with the cognitive actualization of rational potentiality. From the third domain, the affective, philosophical issues converge upon performance testing.

These issues, raised by humanistic and existential perspectives, center on the conflict between external controls or stimuli, pressed upon learners from without to modify behavior and habituate performance, and the free inner choices of autonomous individuals. Furthermore, these issues focus on the legitimacy of criteria for the evaluation of learning and performance. For the sake of economy, efficiency, and social expectations, can standardized, uniform criteria be applied equitably through performance tests to evaluate unique individuals and the worth of their novel abilities, achievements, contributions, and potentialities?

In an even deeper sense, the issues emerging from concerns with the affective domain for the unique worth of human individuality raise fundamental philosophical questions regarding the nature of reality and the sources of truth and goodness. Are human beings essentially, naturally social beings whose originality and uniqueness emerge through varieties of social experience? If so, we may legitimize some social expectations and cultural norms as criteria for individual development. But the primacy of individual subjectivity over social expectations and external standards continues to be philosophically affirmed. And, to the extent that such philosophical arguments possess some admissibility, standardized performance testing will be questioned, and the objective criteria for evaluating performance will be challenged.

John Dewey advocated learning through individual participation in group social problem-solving activities using scientific inquiry and experimentation. This pragmatic approach is based, theoretically, on the interaction between the individual and the sociophysical environment. Thus,

for Dewey and the pragmatists there is no ultimate separation of reality into the subjective and the objective, and therefore, presumably, no inevitable contrary claims of individual subjectivity and external social expectations. But traditional patterns of thought still hold stronger, sway over most contemporary education, and the bifurcation of reality into two competing realms continues to dominate approaches to testing and curriculum design.

For one thing, science in the twentieth century has not been utilized as the intellectual and democratic means to achieve the human community envisioned by liberal pragmatists. Instead, contemporary experimental science has become the handmaiden of technology. In such master-slave relationships among disciplines and cultural forces, free inquiry, intellectual development and social reform usually suffer the consequences of unchecked self-serving interests. Thus, today science is put to the services of many technological projects whose likely consequences may be detrimental to long-range human interests.

Furthermore, the scientific community has not opened to the masses of contemporary society. Even if we do benefit economically or militarily from the technological applications of scientific advances, on the whole, we are generally excluded from the inquiry and experimentation and have little say in the social uses to which scientific discoveries will be put. Therefore, scientific inquiry, as advocated by those who reject the dichotomies of the subjective and the objective, of the individual and the social, has not yet emerged as the means of participating in and contributing to, the direction of human affairs.

The broad cultural consequence for education and evaluation is that we live in a modern, technologized, industrialized world with loyalties, beliefs, and values characteristic of premodern modes of thought. We live daily amid the external securities and conveniences of creature comforts produced and serviced, sometimes efficiently and competently, by technologies and bureaucracies that fragment, dehumanize, and alienate. Yet we also still feel some worth for ourselves and for our humanity, despite our strong dependencies on institutions, systems, and gadgets that we may know how to manage but doubt we can control.

The performance testing movement, also, will struggle with these conflicts and doubts. Can humans be treated and tested merely as reactive objects whose performance is produced and evaluated from without? Contrariwise, how can performance testing serve the interests of unique, purposive learners who creatively choose their own competencies and the qualities and social uses of those competencies? Has performance testing already succumbed to the prescriptions and reductionism of narrow scientism that seeks only to condition and control the predeterminants of performance? Or rather can performance testing be complemented by introspective self-analysis and self-evaluation of individual intentions, plans, volition, and purpose? Will teaching directed toward performance testing facilitate the individual imagination and creativity necessary to construct novel understanding and appreciation of quality performance? In other words, will the performance be taught and tested in such ways that it will serve the needs and interests of the learner, or must the learner serve the inflexible demands of the tests?

Ultimately, these humanistic concerns challenge the functions and uses of performance testing to recognize that those skilled performances are not just economically rewarding and efficient. The performances most worth performing also serve the psychological renewal and self-actualization of the individual.

The Sacrifice of Reality

The problem of distinguishing reality from the mere appearance of reality is as old as philosophic inquiry itself. Some philosophers have argued that what merely appears and is fleetingly perceived is only temporary and thus unreal, not be to be confused with the enduring reality of the underlying form. Other philosophers have defined the very essence of reality in terms of what is perceived, while still others conclude that what is, what exists, cannot be known at all as it is, in and of itself, but rather only as an object of knowledge complying with the categories of human understanding. As such, the question of the nature of reality may raise little interest except among philosophers who value disinterested inquiry into esoteric and irresolvable problems.

And yet the problems of reality and its theoretical distinction from appearance constantly show up in practical, everyday situations, especially in regard to public policy issues in education and evaluation. Performance testing is no exception.

The evaluation of performance is a costly and time-consuming enterprise. Thus, it becomes a practical matter to attempt to reproduce the reality of a job situation through laboratory simulation.

> "While most developers of performance tests strive to retain an element of reality by creating work samples or simulators, there are times when reality must be sacrificed in the interest of efficiency or in the interest of measuring certain mental processes that cannot be measured conveniently in any other way . . . They are quick and easy to use, they do represent important elements of the troubleshooting task, and they can be used in locations where the real equipment cannot. They suffer from their representing only part of the total real environment."[6]

One might add that simulators also suffer from the uncertainty of how well, or to what extent those parts of the real environment are actually represented in the simulated environment.

And no matter how "realistic" simulation appears in performance testing, the performer being tested may have the notion that, except in terms of the evaluation results, the simulation itself "really doesn't count." Efforts to research this problem empirically or experimentally face the difficulty of gathering data and controlling variables of appearance or perception rather than of reality and actuality. Thus, one could never know whether or to what extent the notion of unreality in simulation contributes to or distracts from quality performance. In either case, nevertheless, the reliability of performance tests relying on simulation suffers some unknown degree of distortion due to the "sacrifice of reality." If the performance within a simulated environment does not matter entirely in reality, the performer may be either less cautious or more relaxed, resulting in either better or worse performance.

This, of course, is certainly no devastating argument against simulation and simulators. No one would want to fly in an airplane whose pilot had been licensed only on the basis of pencil-and-paper tests that examine knowledge about technical data. Nor would any of us want to be operated on by a surgeon who had never before used a scalpel. Still, the inevitable divergencies from reality in performance testing should serve as warnings of limitations and reservations. Just as the experimental scientist recognizes that data only approach, never achieve, accuracy, and that the findings are merely probable, tentative, and relative, so also evaluations resulting from the more or less accurate (or inaccurate) measurements of performance in simulated reality cannot be absolutely conclusive, and should not be acted upon or applied as such. Consequently, assessments should be made through a variety of performance

tests other than those using simulation, and through means of evaluation other than performance tests. Again, the argument strengthens the contention that the measurement of manipulative skills alone, to the neglect of intellectual and human relations skills, jeopardizes the entire process of evaluation.

Broader Horizons

The philosophical issues of reality in performance testing expand into ironic complexity. The evaluation criteria in performance testing take the form of behavioral objectives derived from the process of analyzing actual on-the-job skills. One is successful in performance tests to the extent that competencies in job-related skills can be demonstrated, that is, to the extent that the behavioral objectives of vocational or technical training have been achieved. The trainee is held accountable in terms of these behavioral objectives. If the extent of demonstrated skill proficiency is adequate to some agreed-upon standard, then the trainee is licensed, awarded a credential or awarded a certificate or diploma, and hired or promoted and otherwise rewarded for levels of proficiency achieved.

It is a well-known, but slightly understood, fact that from analysis to job-related skills, to the definition of behavioral objectives, to the design of competency-based curricula, to the testing of performance and, finally, to accountability or certification, this training/evaluation scheme locates its fundamental theoretical roots in behaviorism. For behaviorists, all behavior is reactive, a response to stimulation from the environment. And all learning is a conditioned response to external stimuli. Reality consists of external contingencies and observable behavioral responses to them. Therefore, behavior, including competent performance, argue the behaviorists, can be conditioned, controlled, and predicted by managing the environmental stimuli.

It is not the purpose here to provide a definitive critique of the behavioral theory of learning or behavioral technology. It is sufficient to emphasize the behaviorists' reliance on a concept of reality as external and objective, independent of inner mental states and subjective psychic processes that cannot be observed or measured.

The ironic point is that those educational endeavors reliant upon behavioral theory and technology, such as management and accountability by behavioral objectives, including performance testing, cannot afford to surrender the reality from which stimulation, control, and the criteria for evaluation all arise. More specifically, the behavioral techniques utilized in training and testing for competency cannot have it both ways. They cannot exclude from reality, or at least serious consideration, all that cannot be observed and measured, and at the same time for the sake of convenience, efficiency and economy, sacrifice even in part the external reality that is all that remains.

The argument is not that behaviorism and performance testing are wrong in the sense that the theory does not work in practice. Each of us, as a matter of common sense, is only too well aware that our behavior is automatically reactive to external stimuli, and that learned behavior can be uncritically responsive to social conditioning and external reward. We are even gratified that this level of learning through operant conditioning is possible. There is no time for speculative or critical thought when it is past time to slam on the brakes.

Life would be wholly unmanageable if we did not perform most routine and repetitive tasks automatically, without forethought and reflection. Otherwise, we would have to learn and relearn trivia constantly. Survival would then be impossible; or even if it were possible, we would have no

time to reflect upon the reasons, purposes, goals, values, and meanings of surviving in the first place. If behaviors could not be conditioned by responses to external realities, many handicapped and retarded persons could not perform the many tasks of living that most of us take for granted every moment.

Therefore, the argument is not that we cannot, or even that we should not, train and learn and measure behaviors or performances in accordance with the science and technology of behavorism. Rather the point is that the theory underlying the concepts and practices leading up to, and following from, performance testing is inadequate to the degree that it must trade off part of the authentic external reality fitting the job scene for another that, by comparison, only simulates or approximates the appearance of the original.

The answer to this theoretical, if not ethical, dilemma is, of course, not to give up competency-based instruction and performance testing. To do so would render our society and economy totally unmanageable. Instead of giving up the behavior-oriented aspects of competency training and testing, these could be opened up to yet broader aspects and methods of human development, education and assessment not covered by behavioral technology.

For example, humanistic and existential concepts of human nature and behavior, involving free choice, self-direction, and self-evaluation, might be brought to fore. In performance testing, at least, this broader approach requires that the performer be in control in the sense that he has made a deliberate and critically intelligent choice to be evaluated on the basis of a clear comprehension of the tasks to be performed and the criteria to be met. Performance would be viewed and valued as that of a human being with feelings, aspirations, and worth not wholly circumscribed by that performance. In addition to behavioral competencies, human relation and affective skills would be encouraged and rewarded along with critical, reflective intelligence and aesthetic appreciation. The individual skilled worker then is not easily exploited by mass corporate systems, and human life takes on meanings that extend beyond technical proficiencies and occupational settings.

In these broader terms not limited to the independent realities of external stimuli, but including a sense of individual self-worth and pride in proficiency, performers are not subject to impositions that they themselves cannot evaluate, control, and redirect. Their own reality is not reduced to a series of automatic reactions to impersonal conditions and relationships. Performance becomes a way of expressing, realizing, and becoming one's own truer chosen self—not a demonstration of one's ability to meet the expectations, achieve the requirements, or acquire the rewards of others.

Performing Individuals and Individual Performance

It may be that those who strongly advocate performance testing, and especially those who do so uncritically, do so because they discern the performance of the person in the same sense as the performance of a machine designed to operate in some specific fashion. Certainly such a propensity to equate various meanings of "performance" could be understood, if not predicted, especially among vocational and technical educators and occupational evaluators who work with machines, teach individuals to use machines, and test individuals' operations of machines.

If one would not have such expectations or make such predictions of artists, it is not because the artists are better than the vocationalists. Indeed, the two may be one. But, each approaches performance with a different mentality and a different set of presumptions. Workers

use their tools and machines to produce a product or to provide a service; artists use their instruments or mediums to express and create feelings, to interpret and convey meanings and intensions, to provide pleasure and to enjoy the performance.

In the world of work, it is a small but significant step from machine to machinist, to view the performance of the machinist as an extension of the function of the machine. In this sense, the machinist merely completes the otherwise incomplete machine. Thus, the performance of the machinist would be seen as being of the same class as the performance of the machine.

It is this mechanical sense of "performance" that underlies performance or competency-based education. But the performance of a teacher or a worker--of any person--is not the same as the performance of a machine unless one makes no conceptual distinction between persons and machines. Then, and only then, could their respective performances be considered identical.

The performance of a machine must accord with the design of its own production. The sense of an individual's performance "applies to any action of a person who has parts he makes answer to the parts of the work performed, and connects in ways that correspond to relations of the parts of the work."[7] Furthermore, the performance of a person differs from the performance of a machine in that the former depends on the intention of the performer to engage in it. Since this distinction between the performance of a human and that of a machine depends on a theory of human nature as intentional, it somewhat begs the question and is certainly in no sense conclusive. Nevertheless, it is just enough to warn against equating mechanical performance with human performance and thereby applying the same criteria to the evaluation of each.

If work performance cannot be taken for granted as mechanical action, that is, as uncritical application of rules or habits, then at least the theoretical foundations of performance testing are thin and scarce. The performances of machines are not valued intrinsically in and of and for themselves. Mechanical performances are rather valued for their convenient and efficient instrumental functions. Their values lie in their instrumental uses for our own human purposes. What is valuable in human performances does not entirely, at least, depend on this instrumental relationship to our own human interest, or rather cannot do so without rejecting the inherent worth of the individual. It does little good to argue for the inherent worth and dignity of the individual performing and the instrumental value of individual performance. Immeasurable injustice and suffering are historically rationalized by separating the person from the performance, granting intrinsic worth to the individual and mere instrumental worth to the person's "mechanical" performance. Human performance interprets and expresses, some would argue, not only the work patterns and products, but more importantly the meanings, purposes and intentions of the person who, contrary to popular contemporary behavioral technology, cannot, or at least should not, be reduced to a repetoire of measurable, controllable, predictable behaviors.

Relationships of Parts and Wholes

One clear, but problematic, assumption underlying performance testing is that the practice of an occupation is the sum of the tasks into which that occupation has been analyzed and, further, that competency in the vocation can be achieved by learning separately to perform the individual tasks, regardless of their number or nature. Within this assumption, the performance task that is tested is to the vocation as a part is to its whole.

Now the relationship among parts, and in turn their relationships to their whole, may appear at first glance to be simple and straightforward. In some cases, such as with the legs of a chair

and the chair itself, the relationships may be comparatively uncomplicated, though a designer or manufacturer of chairs may argue otherwise. The relationships among human behaviors, and especially behaviors relating several or many humans within a system such as a school or job or entire vocation, can become complex and complicated beyond the point of mere description, much less analysis.

Extending beyond the empirical description, philosophic inquiry has ever been intrigued and challenged by the question of complicated relationships of parts to wholes. In logic, it is fallacious to argue that the qualities of the parts also characterize the whole, or conversely, that the nature of the whole characterizes each individual part. In experience, this may or may not be the case but, if so, never by logical necessity. Of course, philosophy is notorious for its conflicting perspectives, so it comes as no surprise that some philosophic theories prize unity among parts and within wholes, while other pluralistic notions perceive incongruities, if not conflicts, among at least some relationships. Monistic perspectives of unified reality value order, continuity, regularity, and lawfulness among human behaviors and social relationships. Others argue for at least the possibility, if not the desirability, of the diverse, the spontaneous, the innovative, the creative, and the unpredictable.

There is no reason to assume that those engaged in performance-based instruction and performance testing intend deliberately to enter this metaphysical squabble. On the contrary, vocational educators use these training and testing techniques for quite pragmatic reasons that go far beyond or never approaching the desire to argue, even discuss, a metaphysical notion regarding the relationship of parts to wholes, or a social theory advocating the inevitability or desirability of regularity and structure over spontaneity and innovation, or vice versa.

Nevertheless, it must be acknowledged that performance testing and the educational movements on which it depends and with which it is associated are themselves inexorably related to political and educational policies that at least represent, if they do not promote, controversial social values, conflicting educational philosophies, and competing lifestyles.

The performance tester cannot but endorse, or at least sanction, those social perspectives and values inherent within the view that parts relate, or ought to relate, in a unified manner to the wholes to which they rightfully belong; that is, that task performances go to make up the job, or that vocations are the sum of their respective individual tasks. Thus, regularity, predictable performance, consistent production, ordered sequence, dependable service, formal relationships, structured experiences, conditioned responses, reliable competence—these and other similar characterizations make up the reality of human experiences and social relationships observed, measured and monitored by performance testing. No arguments are here proposed against these qualities and processes intensely scrutinized, promoted, and rewarded through performance testing.

But it is necessary to question the degree to which these kinds of values, realities, and beliefs encompass the entire range of human experience and characterize the possible scope of human relationships. When asked, one may be tempted to respond: very slightly. But, even if the predictable qualities and structured processes measured by performance tests characterize most human experiences and relationships, one could again ask critically: Are these ordered sequences and conditioned responses the best parts of the whole sweep of human potentiality? Well, probably not, nor were the elements tested in performance ever proposed to be the highest, most challenging and valuable aspects of humanity—though they may promote higher potentialities, whatever our priorities may perceive them to be.

So, again, a consideration of the philosophical issues in performance testing leads not to the question of whether there is a legitimate, justifiable place for performance testing, but just what is that place in the broader scheme of education, human development and social interaction. Whether one's value orientation or philosophical perspective assigns a relatively high or low priority to the routinized behavioral regularities evaluated through performance testing, it would be as difficult to judge them the best, the finest, the highest as to judge them the worst, the least, the lowest. And somewhere between these two extremes, the measurable performance and the measuring performance test lie as instrumentalities, mere means to competency, social usefulness, and economic independence, but nevertheless as means to yet higher goals of human development and relationships.

Performance testing, like any other means, may be elevated, even for the noblest reasons, to an end in itself. Perceived as such, performance testing no longer serves but defines human existence and experience. That life is likely to be void of diversity and dissent, of innovation and inquisitiveness, of spontaneity and sparkle. It is hoped that the alternatives will not be reduced to a choice between competency, competition, and control on one hand, and creativity, compassion and curiosity on the other. Just as we cannot learn in a rat maze all that is most worth knowing, performance testing cannot evaluate all that we know and are, or should most desire to learn and become.

Conclusion

Performance testing is more than a fad—a mere temporary stop-gap measure for overwhelming perplexities that have been accumulating since World War II. Among those perplexities were: rapidly expanding school enrollments, frantic responses to Sputnik, and charges that our schools were failing, then mobilization to integrate minorities and handicapped persons, followed quickly by social demands for greater equality of opportunity and the need to move from an expanding economy to a steady state. Perhaps at no other time in history has any social institution been called upon to accomplish so much as the American school system in the past generation.

Normally schools reflect and follow the trends of the broader society. Yet, in the past generation, when social goals have been unclear, educators have been called upon to mark out new paths that the broader society has, in many cases been reluctant to travel: integration, conservation, innovation, accountability, economy, reconstruction of traditional belief patterns and value systems. Performance testing, performance contracting, and competency-based instruction are but a few of the major efforts within education to respond without clear social goals or firm social support. No single one of these efforts, or even a combination of several, could meet all the conflicting demands and competing needs placed upon the schools.

A few of the issues raised by performance testing have been reviewed. Its underlying assumptions appear to conflict with both traditional cognitive aims and innovative affective emphases. It raises questions of priority regarding individual autonomy and social responsibility. It appears to contrast the mechanical with the humanistic, the quantitative with the qualitative, the predetermined with the free and open and unpredictable. Performance testing, in fact, brings to the fore the biting theoretical issues and value conflicts plaguing education and our broader society today. As such, it provides a living laboratory for social and educational experimentation.

Experimentation demands caution and control, as well as creativity and courage. Performance testing as an experimental arena is no panacea for all educational problems. Its interests and capacities do not reach all human concerns. The conceptual framework of

performance testing is narrow and shallow compared to the breadth and depth of human prospects and social needs. Its concept of performance is necessarily definite and precise, and therefore not wholly adequate to cover the spectrum of individual interest, will, need and aspiration. Nevertheless, performance testing has its legitimate uses within its defined limitations. The danger is that these limitations will be exceeded when it is called upon to provide more than it has to offer.

Notes

[1]Thomas F. Green, "Minimal Educational Standards: A Systematic Perspective," mimeographed, CEMREL, (September, October, 1977), pp. 3-20.

[2]Ibid., p. 17.

[3]Robert Hutchins, *The Learning Society* (New York: The New American Library, 1968), p. 115.

[4]Ibid., pp. 36-37.

[5]Ibid., pp. 124-5.

[6]Joseph L. Boyd, *Handbook of Performance Testing: A Practical Guide for Test Makers* (Princeton, N.J.: Educational Testing Service, January, 1971), pp. 13, 75.

[7]Kingsley Price, "The Sense of 'Performance' and Its Point," Philosophy of Education 1974: Proceedings of the Thirteenth Annual Meeting of the Philosophy of Education Society (Philosophy of Education Society, 1974), p. 21.

Comments on the Philosophical Issues in Performance Testing

John F. Thompson
University of Wisconsin
Madison, Wisconsin

The two authors present very different ideas. Borow helps the reader learn *about* performance testing while Willers helps the reader learn *of* performance testing. These distinctions are not minor. In learning about something we learn what it is and how it functions. In learning of something we engage in new ways of thinking. It requires us to actively engage in the examination of our assumptions.

Borow helps us understand the history of performance testing, some of its issues and problems. Willers, on the other hand, takes us to basic assumptions and points out inconsistencies with broader goals. The former, then, is more a technical paper and the latter a more philosophical paper. While their differences are sharp and clear they do complement each other.

If philosophical inquiry helps us examine assumptions, what is an assumption? An assumption is something which is taken for granted or supposed and, therefore, cannot be verified in a scientific sense. If an idea can be proved, it ceases to be an assumption and becomes a fact. All of us act on our assumptions—even those that are not examined.

Assumptions need to be examined in light of reliability. A belief is reliable when it always results in the same outcome. Assumptions need to be examined in light of their validity. A belief is valid when it conforms to new knowledge and experiences. And finally, assumptions need to be examined in light of consistency. That is, the entire set of assumptions about a concept like performance testing needs to support and work together rather than against each other.

With this framework in mind, let us examine the two papers. The strength of Borow's paper, I have already indicated, is that it identifies some of the issues and problems of performance testing. Its weakness as a philosophical paper is that it does not go far enough in examining many of the assumptions identified or implied. [illegible], I find the early sections of the paper to be more profound than the latter. Early in the paper the author presents the "tacit assumptions of testing." These are said to be individual differences in people that can be measured, the stability of measured individual differences, and our ability to predict student performance from a test situation to an external nontest setting such as a job. These are powerful assertions. While all are not examined here they need to be by those who favor performance testing.

I admire, particularly, the section on validity. There the author analyzes the assumption of predictability from school to job. It is pointed out that

> "the failure of the typical performance test to tap relevant factors in on-the-job training behavior or bona fide job behavior may limit its capacity to furnish a comprehensive and accurate index of the student's competency. Performance tests customarily

> appraise an array of cognitive and psychomotor skills. Yet, the affective domain is clearly part of the universe of performance on the job. Successful performance in the vast majority of occupations rests at least partially on the worker's attitudes and personal disposition toward the work scene, such as pride of workmanship, compliance with rules of the workplace, quality of interpersonal relations, dependability, and integrity."

The section ends, however, with a practical and technical emphasis on how the validity problem may be solved.

Willer's paper identifies and examines basic assumptions of performance testing. It was necessary for me to read the introduction a couple of times before understanding its purpose, which I concluded to be one of sensitizing the reader to the broad issues. While I wanted to argue with minor points, its conclusion is the focal point.

> "In general, these issues speak to the limitations, narrowness and inadequacies of performance testing when over-stressed or under to the exclusion of other claims and interests. Though threatening to narrow self-interests, such critical messages of caution can provide clarity and breath of purpose together with insights into worthy means of judging human development and achievement."

It is pointed out that:

> "A performance test is presumed to be a measure of occupational competency or the the ability to perform job-related skills. This presumption, in turn, is based on the assumption that job skills, and even overall occupational functions, can be reduced by meticulous analysis to meaningful, manageable and measurable sequential segments. The component performance of these work segments, may be examined and evaluated. The purpose of performance testing, accordingly, is to discern the quality of a particular job-related skill or to qualify for a particular occupation."

I think another assumption needs to be added to this section. We tend to assume in vocational education that if we know which skills are necessary for occupational competence, we know how to teach them. This leads to another dimension that is neglected in this paper. It relates to the lack of assertions about learning theory as it is used to support performance testing.

The remainder of Willers paper is very powerful. It is a very concise philosophical treatise of performance testing. In fact, I wish I had written it.

In sum, while the papers are very different, there are points on which the authors tend to agree. Remember, Borow's paper is more technical. It tends to offer the position that performance testing is rather value-free and its real problems are test validity and reliability. On the other hand, Willers tends to identify assumptions for critical judgments. Nevertheless, they tend to agree that:

- Performance testing is a narrow educational perspective.
- Performance testing does not adequately assess the impact of the affective domain on successful job performance.
- Performance testing has a national perspective.
- Performance testing has a social perspective.
- Performance testing has an inherent conflict between individual and social goals.

CHAPTER THREE

TECHNICAL ISSUES

The technical issues affecting performance testing are either addressed directly or alluded to be every contributor in this handbook. While technical issues such as validity and reliability do cross all of the other issue areas, they have enough importance to stand on their own and warrant a chapter solely devoted to them. Therefore, Chapter Three begins with a discussion of technical considerations by Evelyn Perloff where validity, reliability, efficiency, test bias, and observer/rater variability are addressed. In discussing each of these considerations, she relates their role in classical measurement theory and the applicability of the concepts to performance testing. For example, consistency validity is described as a promising validation approach for performance tests.

Raymond Klein authored the second paper which provides a more pragmatic approach to performance testing. He focuses on developing of performance tests; testing process; standardization and norms; determining cut-off scores, providing test related materials; and revising tests. The chapter concludes with Samuel Livingston providing a third perspective on the technical issues facing performance testing in the Comments paper

Technical Considerations: Validity, Reliability, Efficiency, and Observer/Rater Variability

Evelyn Perloff
University of Pittsburgh
Pittsburgh, Pennsylvania

The purpose of this paper is to describe characteristics of effective testing instruments. There are three crucial characteristics of a good test: validity, reliability, and efficiency. That is, a good test should (1) provide information relevant to announced objectives or uses to which the test will be put (validity), (2) indicate consistent information about those tested (reliability), and (3) be convenient, pertinent, and economical to administer and interpret (efficiency). It is generally conceded by measurement experts that the most fundamental characteristic of a good test is validity, with reliability generally considered secondary. Least important are those additional considerations that include efficiency and a variety of characteristics which reflect a test's utility.

This paper discusses characteristics of a good test that derive from classical measurement theory. Although performance testing calls for modifications of classical measurement theory, these revisions have been slow in coming and as a result much of classical measurement theory remains appropriate. There are, however, some hopeful indications that useful changes are being developed for specific evaluation of performance tests. These will be presented here whenever appropriate. Two procedures of particular concern to performance testing are observing and rating what individuals do in test situations. The last section of this chapter will therefore present a brief consideration of both procedures, with special attention to the issue of observer and rater variability.

Validity

Although validity is considerd the most important feature of measuring instruments, it remains the most difficult to assess because it is the most complex. Furthermore, validity involves a number of considerations that are external to the test itself, yet need to be related to test performance. Validity has been defined in several ways, but these definitions stress the same general idea: Does the test measure what it is supposed to measure? If the answer is yes, then the test is considered valid, if the answer is no, then the test is not regarded as valid. Validity is a matter of degree not an "all or none" condition. That is, two tests can be assessed as valid, but one may be more valid than the other because it does a better job of measuring what it is supposed to measure. There are also four different kinds of validity. Depending on how validity is defined, the four kinds are: (1) criterion validity, (2) content validity, (3) construct validity, and (4) consistency validity. The first three apply to classical measurement testing and the fourth is more specific to performance testing. The four kinds of validity are discussed below.

Criterion Validity

Teachers and managers frequently need to compare test achievement with school or job performance. That is, tests are administered because it is necessary to predict present or future abilities. The emphasis here is not on what the test measures, but rather on how well it predicts, that is, the quality of the test is not determined by the test's content per se, but rather with the ability of performance of that content to predict later school achievement or job performance. If subsequent school or job expectations, based on earlier test performance, are confirmed then the test has criterion validity. Criterion validity has also been called criterion-related or concurrent and predictive validity.

Criterion validity is so termed because it relates to a criterion (standard) or rule for judging the value of something. In measurement, a criterion is performance (academic grades, supervisory ratings, job proficiency) against which the value of a test score is judged. Thus, a test has criterion validity if individuals who are judged successful on the criterion (do well in school, obtain high job ratings, perform effectively on-the-job) are those who also obtained the high test scores. Similarly, we would expect individuals who are judged unsuccessful on the criterion (do poorly in school, obtain low job ratings, perform inadequately on-the-job) to be those who obtained the low test scores. In contrast, a test does not possess criterion validity if there is little agreement in how individuals perform on the criterion and how the test assesses their abilities. That is, higher test scores correspond to a range (low and high) of school or job measures and low test scores correspond to a range (low and high) of school or job measures.

Criterion validity presupposes that a criterion is relevant and has been accurately measured. That is, not any criterion will do. A criterion must be salient for those who wish to make personnel decisions on the basis of test scores. For example, grades are viewed as a salient school criterion, but number of hours studied or ability to outline material effectively, although worthwhile and perhaps means to an end for grades, may not in themselves be considered good criterion measures. Obviously, selecting a criterion is no easy task since the complex and difficult issues inherent in the concept of validity are true for criteria as well as for tests. This predicament is readily observed for the two most commonly used (and supposedly most appropriate) criteria: school grades and on-the-job performance ratings. Unfortunately, too little effort is expended on the criterion side of the ledger. We suspect that until this state of affairs is modified, criterion validity may not accurately reflect an instrument's effectiveness.

Content Validity

Judging the adequacy of a test's substance or content describes the process of content validation. It seeks to answer the question: Does the test measure what the test constructor (teacher or manager) thinks it does? Judgment in this context generally refers to evaluation by experts in a content area.

Content validity is typically applied to tests measuring outcomes of education and training. For the most part, these tests are achievement tests or representative samples of the universe of appropriate content. The process of content validation specifies clearly defined steps to ensure that the final product, the test, has maximum content validity. A first step involves relating instructional content on the one hand to a taxonomy of objectives on the other. This step encourages delineation of expected instructional outcomes as well as detailed student behaviors. It resembles preparation of an efficient lesson plan.

Following this, appropriate measures of the expected instructional outcomes and student behavior can be developed. This second step requires representative sampling of the test's content. The completed test is then ready to be judged by appropriate experts in the area regarding adequate coverage of its content. If the experts concur, the test is considered to have content validity. If the experts are critical and disagree, the test will not be assessed as having content validity.

Although content validation can progress in an orderly fashion, execution of the primary steps—subject matter selection, outcome specification, content sampling, and ultimate judgment by experts—tends to be highly subjective. It appears unrealistic, then, to expect constant close correspondence between what a test author includes in a test and how that test is judged by experts in the field. Unfortunately, in many situations, there may be no other alternative than content validity as a measure of a test's effectiveness. Lennon states it well when he says that

> in many testing situations (of which achievement testing forms the largest class) there is not available or readily accessible any dependable criterion variable, against which the "validity" of the test may be measured; and secondly, is the fact that there are certain uses of tests for which correlations with either contemporary or subsequent criteria are not meaningful as indicators of validity.[1]

It is probably with regard to content validity that performance tests fare best. Their contents appear to resemble the objectives and contents established by a curriculum and are therefore readily acceptable to educators and job trainers. In fact, Borow points out that when performance tests have

> highly relevant content they are so compellingly convincing in appearance that vocational educators, on-the-job (OJT or JIT) training supervisors, and industrial personnel recruitment officers are tempted to accept derived scores from such performance tests as tantamount to job proficiency.[2]

A final issue regarding content validity that pertains specifically to performance tests as they relate to minimum competency testing is to view content validity in terms of curricular and instructional validity.[3] Curricular validity determines how well a test measures a curriculum's objectives. This involves a comparison of test and curriculum objectives. Instructional validity measures whether the schools provided the content assessed by the test. Both curricular and instructional validity place additional burdens on tests that are beyond that generally demanded by content validity in measuring student and employee performance.

Construct Validity

Whenever it is necessary to consider one or more underlying properties or constructs (concepts) that an instrument measures, then the relevant validation procedure called for is construct validity. It is an analysis of the meaning of test scores in terms of psychological concepts or "constructs". This kind of validity is considered the most significant and important because it derives directly from theory. Unlike criterion and content validities, the process of construct validity is not easy to understand. It is intricately linked to science and is the same process as that used to generate and test scientific theories.

There are various types of evidence that can be considered in establishing construct validity. The term "construct" refers to an underlying trait, disposition, or ability, such as anxiety, congeniality, motivation, responsibility, or verbal influence. These are five examples of a very large number of possible constructs. There are two primary ways to obtain evidence of construct validity. First, a test has construct validity if it differentiates between individuals who rank high and those who rank low on the construct underlying the test. (Note that in this case the construct is in fact a criterion measure.) Second, a test has construct validity if the theory proposes certain modifications of the construct and these in turn produce corresponding test score changes. Most frequently construct validity is accomplished by examining a group of tests believed to be measuring the same thing. Then, the characteristic underlying what is common to these tests (a construct) is identified by using a statistical procedure called factor analysis. The technique of factor analysis permits reduction of a complex domain of many variables to one of simpler structure with fewer variables. This analytic procedure identifies tests or measures that are closely related (highly correlated) with one another. That is, these tests or measures are similar, they belong together. The reduced number of characteristics or variables underlying groups of similar tests or measures are then called factors or constructs. It is important to remember that the construct identified will depend on the specific tests and measures included in the factor analysis. According to Ekstrom,[4] there are a number of problems affecting construct validity when factors of a factor analysis are used as criteria. The first problem results when a number of tests identified by the same construct are actually measuring different things. Second, characteristics of examinees affect the factor structure of a test. That is, definitions of mental health differ by sex and race. For example, if males and females exhibit the same behavior, it "may be rated as highly aggressive for the female but only moderately aggressive for the male."[5] Similarly, some personality measures are affected by race "because nonpathological racial variance contributes to elevated scores on some scales."[6]

A third and last concern relates to examinees' use of different strategies to solve problems presented in tests. For example, it has been demonstrated that although many individuals mentally manipulate figures in solving spatial visualization problems, others use an analytic strategy to separate figures into elements and then look for similarities. Similarly, according to Gruen and Parkman,[7] most adults use memory to solve problems of simple addition, but most children and some adults use incremental counting to solve these problems.

Construct validation is obviously a much more complex and time-consuming process than either criterion or content validity. As described by Cronbach[8] "construct validity is established through a long-continued interplay between observation, reasoning, and imagination"; and according to Kerlinger[9] it has been "recognized as a central kind of validity" by the American Psychological Association. In summary, construct validity appears as the most promising validation procedure, worthy of the necessary time, effort, and expense required to identify as well as measure the relevant construct.

Consistency Validity

The previous discussion has presented the classical model used to establish test validity. That is, as pointed out by Wernimont and Campbell,[10] the classic validity model uses tests "as signs, or indicators," instead of sampling appropriate behaviors to predict future performance. As many writers have pointed out, particularly those who have encountered a variety of difficulties in trying to predict on-the-job performance, the classical model has not always been effective. There is substantial evidence to indicate that validities for many predictors (measures of mental ability, specific and general aptitude measures, achievement tests, interests, or personality dimensions) of job performance have remained low. In fact, these conditions have persisted for

over 50 years in spite of extensive efforts by professionals in government, industry, and the military to ameliorate this state of affairs. Hopefully, we are finally ready for "an idea whose time" has long since passed. What is being proposed, then, is to modify criterion validity as it is now defined to stress consistency between criteria and predictors. Or as Wernimont and Campbell state:

> The essence of the suggested procedure is the establishment of consistencies between relevant dimensions of job-behavior and preemployment-behavior samples obtained from real or simulated situations. If samples instead of signs are employed, a number of prediction and measurement problems seem to be alleviated or at least confronted more directly.[11]

In other words, the shift in criterion validity that is being suggested is from predictors as signs to behavior as samples of future performance. Wernimont and Campbell describe it well when they say, "The best indicator of future performance is past performance."[12]

A related issue here is a tendency by those in measurement to refer to any relationship between similar behavioral measures as reliability rather than validity. Classic measurement theory defines validity as the correlation between dissimilar predictors and criteria. In contrast, consistency validity looks to relationships between similar predictors and criteria. This latter notion of behavior sampling appears to be the basis of a large domain of performance assessment: namely, simulation. Wernimont and Campbell also point out that the approach seems to underlie prediction from biographical inventories that include items that "represent an attempt to assess previous achievement on similar types of activities."[13]

Four possible steps constitute application of the consistency model. The first steps entails an extensive job analysis, with specific attention to those job dimensions which relate to critical behaviors for successful and/or unsuccessful job performance. Second, each applicant's background (education and experience) is assessed for manifest critical behaviors. Step 3 follows whenever an applicant's background data do not include relevant job behaviors. This step requires administration of numerous work-sample tests and/or simulation activities. The fourth and final step involves use of "individual performance measures of psychological variables"[14] whenever possible.

A final issue involved in consistency validity is that predicted measures must not only be behavioral measures but also observable job behaviors that relate to performance competency. Behavioral measures of the performance of, say, a production manager would refer to assessment of such job activities as scheduling requirements, operating costs, spillage and waste, employee absenteeism and tardiness, procurement, and future planning. These become the predictor measures (or behavior sample) and must be similar to and therefore predictive of the criterion (or measures to be predicted). It follows then that such frequently adopted criteria as salary increases and promotions are inappropriate. Neither criterion can be considered a job behavior nor can the individual exert significant control over either of them.

Some Advantages. Although consistency validation is not a total panacea for problems associated with criterion validity, it can provide better returns in seeking to understand job performance by stressing behavior measurement. As suggested by Wernimont and Campbell,[15] there are four primary advantages that consistency validity has over criterion validity. These are presented below.

1. *Stability of relevant job behaviors.* In spite of productive research relating to performance criteria there appears to be little information documenting stability of relevant job behaviors. It follows then that consistency validity, which speaks to recurring and relevant job behaviors, is a more applicable validation approach than classical methodology attempts "to generalize from a one time criterion measure to an appreciable time span of job behavior."[16] That is, unlike criterion validity, consistency validity stresses longitudinal prediction.

2. *Faking and response sets.* Since consistency validation maximizes behavior and minimizes self-reporting, the usual response biases that affect self-reports will be significantly reduced.

3. *Discrimination in testing.* As pointed out by Doppelt and Bennett,[17] two common criticisms made against tests are (1) lack of relevance and (2) unfairness of content. Both charges have had deleterious consequences on testing programs, particularly in business and industry. Thus, a number of legal cases have shown that many job skills and knowledge can be obtained through on-the-job training programs, regardless of test performance. Similarly, many tests have been considered "culture-dependent." Test items stress white middle-class values that result in an inaccurate appraisal of the disadvantaged or those who have not been influenced by white middle-class culture or education.

4. *Invasion of privacy.* This is the fourth and final problem that the consistency validity approach dissipates. That is, there is neither need to develop new tests each year nor maintenance of strict security over testing materials by test developers. The tests, by specification and design, are to resemble job behaviors. Thus, these behavior samples, by their very nature, serves as obvious links between preemployment and on-the-job behaviors.

Consistency validity appears to be a promising validation approach. It is suggested as a replacement for criterion validity only (not for construct validity), and it is particularly appropriate for performance tests because it focuses on the measurement of behavior. That is, consistency validity substitutes behavior samples for predispositional signs, stresses longitudinal over one-time criterion measurement, and can significantly reduce persistent testing problems of response sets, discrimination, and invasion of privacy.

Reliability

Reliability, Cronbach perfers the term generalizabilty,[18] is the second most important characteristic to consider in evaluating measuring instruments. A variety of terms have been used to define reliability. They include accuracy agreement, consistency, dependability, generalizability, homogeneity, precision, regularity, stability; and trustworthiness. Of these terms, consistency is probably considered most representative, although not totally encompassing. Consistency here refers to stability or trustworthiness of test performance over time. Unfortunately, reliability measurement involves an indirect and statistical conceptualization. Thus, it is assumed that a "true score" exists on a particular test for every individual, but these scores are indeterminate. They could, however, be approximated if the test were administered many times to the same individual. Not only is this unreasonable but it is also unrealistic since a test is usually administered only once. Hence, reliability is interpreted as that proportion of the variance attributed to variation in the "true sense."

Estimation is essential here because behavior fluctuates, with the result that performance varies from one time to the next. Furthermore, no single measurement can be expected to typify

an individual's behavior completely; it can only serve as a rough approximation. Test theory provides techniques for assessing this variability of test scores in order to estimate "true" performance. The most familiar estimation approach is to determine the standard error of measurement that provides an indication of the magnitude of error between "true" and observed performance.

There are additional approaches for assessing reliability, where comparisons require making at least two observations per person. The emphasis here is on consistency or lack of error. That is, a reliable test is one that is devoid of error, where error refers to test score inconsistencies resulting from a variety of influences and conditions that plague measurement. These errors are random or chance fluctuations that do not result from changes due to the nature of what is being measured, but may result instead from variability on the part of the test taker, due to such factors as fatigue, low or high motivation, and variability in the interpretation of ambiguous test questions.

There are two comparisons for checking consistency: (1) administering equivalent parts or complete tests on the same occasion and (2) administering the same test on several occasions. The former approach (measuring on one occasion) indicates how well two sets of comparable test scores agree when they have been obtained at the same time. The latter approach (measuring on several occasions) compares agreement of two or more sets of test scores when they have been obtained at different times. Both approaches examine the four major sources of test-score variation that affect reliability. These have been succinctly specified by Cronbach[19] as four kinds of charcteristics that influence an individual's performance: (1) lasting and general characteristics, (2) lasting and specific characteristics, (3) temporary and general characteristics, and (4) temporary and specific characteristics.

Measuring on One Occasion

Two methods are available for determining reliability in this situation: alternate form (administering equivalent forms of a test) and internal consistency (dividing a single test into equivalent parts). The two major sources of test-score variation that are counted as error here and hence reduce reliability include both lasting and temporary specific characteristics of the individual. These specific characteristics are appropriately illustrated by (1) lasting skills, abilities, attitudes, and knowledge called for by the particular test, and (2) temporary memory fluctuations, motivational changes, luck, and emotional states related to the particular test.

Measuring on Several Occasions

Two methods are also available for determining reliability over time: retest (administering the same test after an appropriate time interval) and delayed alternative-forms (administering equivalent forms of the test after an appropriate time interval). The exact length of the interval is not of major concern, only that it be long enough to minimize effects of memory. The two major sources of test-score variation that count as error in this case are general and specific temporary characteristics of the individual. These temporary characteristics are fittingly illustrated by (1) general knowledge, skills, attitudes, and habits related to the particular test, and (2) specific memory fluctuations, motivation changes, luck, and emotional states related to the particular test.

Although validity is considered the single most essential requirement of a good test, reliability helps to ensure a test's trustworthiness and dependability. Cronbach sums it up well:

> Information on reliability is supplementary. It sometimes warns us that validity will be limited just because of error of measurement, and it sometimes helps us plan a more accurate data-gathering procedure.[20]

Efficiency

The third characteristic to be desired in tests is efficiency. This characteristic refers to a number of supplementary considerations that include sources of test bias, "face validity," applicability, and cost. Although none of these is as conceptually critical as validity, they do relate to the test's effectiveness and should be examined as part of a test selection procedure.

Sources of Test Bias

As discussed by Ekstrom,[21] tests should be as free as possible from different types of content bias. These biases are (1) numerical, (2) role, (3) status, (4) stereotypic, and (5) familiarity. The first four biases result when members of certain groups are underrepresented or overrepresented by number, level, kind, and stereotype of activities in which they are portrayed in tests. The fifth and final bias, familiarity, results when certain groups have had differential opportunities for experience or familiarization with specific test content.

As Ekstrom[22] points out, these biases have been well documented in the literature. Numerical bias has frequently occurred because women are infrequently presented in achievement tests. In contrast, role bias has been frequently found in test content because women are generally portrayed as housewives, secretaries, and teachers, suggesting that women do not (or cannot) enter all occupations. Similar to role bias is status bias where women and minorities are rarely presented in administrative and leadership positions. That is, they are teachers and salespersons but not principals and managers. Stereotypic bias results when tests show (1) women as less interested or able to work with mechanical equipment, preferring instead to work in homemaking and helping areas; and (2) minorities as less interested or able to handle the professions, preferring instead to remain as blue-collar workers.

The fifth and last bias, familiarity bias, is best illustrated by Ekstrom when she describes a spatial visualization test "in which the subjects were told that the process involved in solving the problems is similar to 'working with sheet metal'. Such a statement probably biased this test in favor of males because it suggested that these items can only be solved by people who have some knowledge of sheet metal work. The identical process could have just as accurately been described as similar to 'working [redacted] dress pattern.'"[23]

It is a sad commentary, indeed, to point out that not only do these biases affect performance, but also that there is little, if any, research data to substantiate or refute them.

Face Validity

This consideration refers to the nontechnical issue of consumer appeal. That is, public acceptance of a test generally demands that it appear relevant and meaningful. A test that

appears appropriate and reasonable to those examined is said to have "face validity." Although no quantitative assessment can be made of face validity, Cronbach[24] states, "if a test is interesting and 'sensible,' taking it is likely to be a pleasant experience," and this probably produces valid scores

Scores in this case pertain to the specific behaviors of the test and as such indicate the test's purposes. It is important to consider two questions: Does the test measure what it is assumed to measure? Does it adequately sample the appropriate content? In most cases, tests measure what they appear to be measuring, but there have been occasions when this has not been so. That is, so-called clerical aptitude tests with subtests seeking to measure numerical, equipment identification, and information abilities have been found to be predictive of mechanical aptitude. Thus, as Selltiz, Wrightsman, and Cook[25] caution: "'just looking' is smug ignorance." Technical validity (as previously presented) should not, of course, be sacrificed for face validity and this is not necessary because tests that have both technical and face validity are usually available.

Applicability

A third measure of efficiency is the ease with which a test can be administered, scored, and interpreted. A test is easy to administer if it does not require highly trained persons to administer it. Similarly, a test that does not have either complex or specifically timed instructions will be easier to admnister. A test that can be objectively scored will be easier to handle than a test that requires judgment of observation. And finally, a test that can be readily interpreted and communicated by prepared check lists or tables is easier than a test that requires professional expertise for interpretability and communicability.

Cost

The last consideration of efficiency is cost of test materials, administration, and scoring. Costs can be reduced when it is possible to reuse test materials. If a large number of individuals are to be tested, it may be more economical to obtain a full-service package from the test publisher that covers test materials, scoring services, and reports of individual and group results.

Summary

In summary, a test is efficient when it is unbiased, acceptable (has face validity), applicable (easy to administer, score, and interpret), and economical. Decisions regarding tests must initially consider relevance and consistency of information. For this assurance, we turn to validity and reliability. A final, but not necessarily insignificant consideration, is test efficiency. Certainly, if validity and reliability of two tests are about the same, the decision regarding which test to use should be based on matters of efficiency.

Observer/Rater Variability

Dicussion thus far has concentrated on issues from classical measurement theory—validity, reliability, and efficiency—that affect testing. As pointed out by Klein,[26] however, a performance test "involves observing and rating what individuals working at specific jobs in a variety of situations and conditions actually do." As a result, developers of performance tests face

particularly difficult problems that do not generally confront those who design the more typical achievement and aptitude tests. We refer here to obstacles inherent in the processes of observing and rating. More specifically, we will limit discussion to validity and reliability issues of observation and rating measures, presenting observation issues first.

Observation Issues

Variability among observers arises primarily from two sources of error: variability within individual observers and a variety of systematic observer biases. As presented by Simon,[27] observer variability means "the inability of a given observer to repeat an observation again and again in exactly the same way with exactly the same result; and bias means "a tendency to observe the phenomenon in a manner that differs from the 'true' observation in some consistent fashion."

Overcoming observer bias is not an easy task. Biases appear to creep in regardless of how much care is exercised. Ideally, then, the task "is to determine each observer's bias and allow for it."[28] Since this is highly unlikely, a more realistic approach is to use a number of tactics specifically developed to decrease variability within observers which, in turn, also reduces variability from bias among observers.

Six common tactics suggested by Simon[29] that have been found helpful include: (1) sufficient training of observers, (2) detailed specification of tasks that observers are asked to perform, (3) provision of specific written instructions for constant consultation by observers, (4) reporting information as soon after observation as possible, (5) use of mechanical devices whenever appropriate, and (6) obtaining information from several observers who observe at the same time. As pointed out by Simon,[30] these tactics "reduce the area of discretion within which bias may operate" by (1) providing carefully detailed protocols for observers to follow, (2) discouraging inferences from observers, and (3) stressing techniques that minimize forgetting and inaccuracy. Thus, Hulett[31] has advised "that a stubby pencil and a small battered notebook make people less nervous than do more pretentious tools."

Observer reliability is concerned with interobserver agreement as well as with the agreement of individual observers over time. It is, however, usually defined as "the degree to which two or more observers agree on their observations."[32] There appears to be no consensus on a single formula to use in determining observer judgments, but a common method is to divide number of agreements by the sum of number of agreements plus number of disagreements.[33]

According to Selltiz, Wrightsman, and Cook,[34] this formula demands a brief observation time to ensure that observers code the same unit of behavior. The formula gives overly high reliability values when percentage agreements are compared with chance levels, and there are there are high base percentages and few categories.

Rating Issues

Performance testing also frequently requires ratings of learning or work activities. For example, to measure learning obtained in a short-term library experience, "we can complete a

rating scale to assess [his/her] learning, using such criteria as relationship to patrons, accuracy of information provided, cooperativeness and attitude."[35]

Unfortunately, a variety of systematic errors are also present in ratings. For the most part, these errors result from rater biases. Three common systematic errors include halo effect, generosity error, and contrast error.

Halo effect results when raters generalize their impressions from one rating to another. That is, they seek to achieve consistency or what Newcomb has called "a 'logical error'; that is, judges often give similar ratings on traits that seem to them to be logically related."[36] Generosity error occurs when raters overestimate positive qualities of individuals whom they like. Similarly, raters appear to judge individuals as belonging to middle categories rather than assigning them to the extremes. According to Murray, contrast error results because of "a tendency on the part of raters to see others as opposite to themselves in a trait."[37]

There are, in addition, a number of sociocognitive biases that can be expected to affect ratings. Thus, raters may

> attach excessive weight to information that is highly concrete,salient, and easy to remember . . . may be prone to overestimate the extent to which behavior is caused by stable personality factors, while minimizing the impact of situational and environmental forces on individual's behavior . . . and, because people are unaware of fundamental statistical principles, they are susceptable to biases in judgment.[38]

These biases include only a portion of those that can influence judgment. Both validity and reliability are reduced not only by systematic and random errors, but also by the many sociocognitive biases that may occur. Unreliability of ratings among raters frequently results from "the fact that some frame of reference is implicit in any rating; different raters may use different frames of reference in describing individuals in terms of the characteristics in question."[39]

It is fortunate therefore that a variety of ways exist for reducing errors and biases. Although it is not possible to list the many techniques for minimizing these influences, we offer several ways to improve rater accuracy, in addition to those listed for overcoming observer variability. For example, one suggestion to reduce the constant errors described previously is not to use extreme rating scale positions such as: The student always uses proper lighting in taking photographs. A preferable (less extreme) statement would be: The student generally uses proper lighting in taking photographs. Similarly, the use of neutral descriptive scale positions instead of evaluative ones are likely to reduce generosity error. Biases can be avoided by adopting a scientific approach and maintaining awareness "of the fallibility of judgmental processes."[40]

Summary

In summary, a technical discussion of performance tests should include issues of observer and rater variability in addition to the classical measurement processes of validity, reliability, and efficiency. Although observer variability is not easy to control, a number of tactics can be adopted to reduce variability within observers which, in turn, also reduces variability among observers. Ratings generally include a variety of systematic errors and sociocognitive biases that affect both validity and reliability. As with observer variability, rating errors and biases can be significantly minimized by adopting a number of similar techniques, the primary one of which

stresses a scientific approach. It is apparent therefore that overcoming biases and errors is difficult and, regardless of how much care is exercised, they appear to creep in. The best solution to these problems seems to be to use the variety of tactics specially developed to decrease variability and increase validity and reliability.

Notes

[1]Roger T. Lennon, "Assumptions Underlying the Use of Content Validity," *Educational and Psychological Measurement*, 1956, p. 297.

[2]Henry Borow, "Philosophical, Practical, and Technical Issues Pertaining to Performance Testing in Vocational Education," (Columbus, Ohio: National Center for Research in Vocational Education, 1979), p. 18.

[3]Stephen J. Slater, "Applied Performance Testing: Typology, Advantages, Limitations, and Examples," (Portland, Oregon: Clearinghouse for Applied Performance Testing, Northwest Regional Educational Laboratory, 1978), p. 26.

[4]Ruth B. Ekstrom, "Issues of Test Bias and Validity," Paper presented at Symposium, Revising the 1974 APA Test Standards, American Psychological Association Annual Meeting, New York City, September 1979), pp. 10-11.

[5]Ibid.

[6]Ibid.

[7]Ibid., p. 12.

[8]Lee J. Cronbach, *Essentials of Psychological Testing* (New York: Harper & Row, 1970), p. 142.

[9]Fred N. Kerlinger, *Behavioral Research: A Conceptual Approach* (New York: Holt, Rinehart and Winston, 1979), p. 141.

[10]Paul F. Wernimont and John P. Campbell, "Signs, Samples, and Criteria," *Journal of Applied Psychology*, 52 (May 1968): pp. 372-76.

[11]Ibid.

[12]Ibid.

[13]Ibid., p. 373.

[14]Ibid., p. 374.

[15]Ibid., p. 375.

[16]Ibid.

[17]John P. Doppelt and George K. Bennett, "Testing Job Applicants from Disadvantaged Groups," *Test Service Bulletin*, No. 57 (New York: Psychological Corporation, 1967), pp. 1-5.

[18]Cronbach, *Essentials of Psychology Testing*, p. 154.

[19]Ibid., p. 175.

[20]Ibid., p. 182

[21]Ekstrom, "Test Bias and Validity," p. 1.

[22]Ibid., pp. 2-3.

[23]Ibid., p. 4.

[24]Cronbach, *Essentials of Psychological Testing*, p. 182.

[25]C. Selltiz, L. S. Wrightsman, and S. W. Cook, *Research Methods in Social Relations*, 3d ed. (New York: Holt, Rinehart & Winston, 1976), p. 179.

[26]Raymond S. Klein, "Some Selected Technical Issues Related to Performance Testing," In *Performance Testing: Issues Facing Vocational Education*, (Columbus, Ohio: National Center for Research in Vocational Education, 1980).

[27]Julian L. Simon, *Basic Research Methods in Social Science: the Art of Empirical Investigation*, 2d ed. (New York: Random House, 1978), p. 273.

[28]Ibid., p. 276.

[29]Ibid., pp. 278-79.

[30]Ibid., p. 278.

[31]J. E. Hulett, "Interviewing in Social Research: Basic Problems of the First Field Trip," *Social Forces* 16 (March 1938): p. 364.

[32]Selltiz, Wrightsman, and Cook, *Research Methods*, p. 287.

[33]Ibid.

[34]Ibid., p. 288.

[35]S. Malak, J.E. Spirer, and B. Land, *Assessing Experiential Learning in Career Education*, (Columbus, Ohio: National Center for Research in Vocational Education, 1979).

[36]Selltiz, Wrightsman, and Cook, *Research Methods*, p. 287.

[37]Ibid.

[38]R. M. Perloff, V. R. Padgett, and T. C. Brock, "Socio-cognitive Biases in the Evaluation Process," In *Ethics, Values, and Standards in Program Evaluation*, edited by Robert Perloff and Evelyn Perloff (San Francisco: Jossey-Bass, in press).

[39]Selltiz, Wrightsman, and Cook, *Research Methods*, p. 409.

[40]Perloff, Padgett, and Brock, "Socio-cognitive Biases in," p. 25.

Some Selected Technical Issues Related to Performance Testing

Raymond Klein
National Occupational Competency
Testing Institute (NOCTI)
Albany, New York

Developing Performance Tests

The current emphasis in performance testing is to develop measures of direct assessment of skill attainment. In order to to this, a candidate is asked to perform a series of tasks based on actual jobs that have been judged critical in relation to demands of a specific occupation. In this context, "critical" means the demonstration of skills considered essential to perform adequately in a specific occupation. In order to be able to construct valid performance tests, the test specialist needs to obtain a timely occupational analysis from which the critical competencies and tasks may be determined. Once these critical competencies have been uncovered, they should be ranked in order of the frequency in which they occur, as well as their relative importance in the job. In this fashion, a list of critical competencies may be identified. Essentially, these key competencies set one role apart from another by identifying the elements that give the occupation its uniqueness. (See Table 1.)

Unlike teacher-prepared examinations that can be put together after identification of the objectives of a unit of instruction, a performance test designed to measure occupational competency requires more extensive efforts to construct.[1] Conducting an occupational analysis involves observing what individuals working at specific jobs in a variety of situations and conditions actually do. Out of this observational data, a categorization of the occupational competencies must be made. The categorization provides the developer with a distribution of a variety of tasks into divisions, each division representing more or less a unique major factor of the particular occupation.

Each major division then has to be broken down into its respective subdivisions, thereby grouping all subtasks into an orderly structure. After the information collected has been so categorized, it should be reviewed by knowledgeable people in the field to confirm the validity of the breakout. Having obtained a measure of consensus from knowledgeable individuals regarding the competencies that comprise specific occupations, it is then necessary to reorganize the specific tasks in a hierarchical manner so that the least critical competency is placed on the bottom of the list and the most sophisticated understanding appears at the top of the list. The competency with the highest point total (frequency X importance) would appear first, the other competencies would be placed in a descending rank order. Once this is accomplished, it is necessary to identify examples of jobs or tasks based on these actual job-related competencies, and to consider them for inclusion in a performance examination.

Table 1

A Model for Determining Actual Competencies

1. Identification of major divisions of the occupation
2. Identification of subdivisions of each major division
3. Identification of competencies required for each subdivision
4. Identification of critical competencies

Critical Competencies = Frequency of Use x Importance to Job

a. **Frequency** may be scaled:

Criteria	*Weighted Value*
Very frequent	(5)
Frequent	(4)
Average	(3)
Occasionally	(2)
Rare	(1)

b. **Importance** may be scaled:

Criteria	*Weighted Value*
Critical	(5)
Essential	(4)
Importance	(3)
Needed	(2)
Desired	(1)

Note: The jobs and tasks selected for inclusion in the performance test should measure an array of the critical competencies both directly and subsumed.

In the past, because of the cost and time required for such undertakings, comprehensive catalogs related to specific occupational competencies were rarely assembled. In recent times, with states pooling their resources, organizations, such as V-TECS and the Ohio Instructional Materials Laboratory, have been compiling catalogs of competencies related to the specific occupations. These ventures have, in turn been translated by various state departments of education into curricula aimed at developing the specific critical skills. These same analyses that are used to identify occupational skills and knowledge are also helpful to test developers for selecting those competencies that need to be assessed by means of a performance examination.

Major Steps

Specific jobs or tasks have to be determined based on the critical competencies that were identified. These jobs or tasks can then be used as the vehicle to assess skills. To develop the test, it is advisable to bring together a committee of practitioners and teachers of the occupation. This committee is used to identify the jobs and tasks that will be required to test a candidate's understanding of the critical competencies needed in the work setting. This can be accomplished by having the committee:

1. Review the specific competencies and then identify potential tasks or jobs.

2. Hypothesize regarding what might be appropriate jobs or tasks and then validate or change the jobs through a process based on the analysis of the occupation. (In practice, both approaches, individually or combined, are used.)

The competencies related to a specific occupation can also be arranged by level. For example, skills usually identified with a skilled worker would be different, in certain respects, from those of an apprentice.

Therefore, the competencies could be categorized by job levels within occupation. Organizing the competencies by level will help the test developer design examinations more appropriate to a specific job or jobs within any occupation. Organizing by level will require the additional breakouts related to major divisions and subdivisions of competencies. These listings need to contain the actual understandings and skills required to function adequately at each level.

In summary, once the major divisions have been identified and the competencies within each level described, specific understandings and skills within each subdivision can be ascertained. Such information forms a basis for curriculum development as well as for the construction of performance tests. Occupations are broken down into specific job levels, and in turn, each level is arranged into specific competencies. The scope of each examination must be specific to the level desired. The jobs selected for inclusion in the test should be based on these levels as well and they should be representative of current practice in the occupation. (See Table 2.)

Additional Considerations

The jobs selected for inclusion on the performance test should evaluate different competencies. Each job should adequately measure specific aspects of the critical competencies required in the performance of the occupation. When a student undertakes to identify what may be causing a malfunction, the logic of the troubleshooting approach should be assessed. There must be a demonstration by the student of approved methods.

Table 2

Major Developmental Steps Related to Constructing Performance Tests

1. Identification of the field and level of jobs within each field
2. Determination of competencies through occupational and task analysis
3. Organization of competencies by level
   a. single skilled
   b. semiskilled
   c. skilled
   d. technical
   e. professional
4. Categorization of competencies by job level
5. Analysis of competencies per job level to identify critical competencies
6. Identification of jobs or tasks by which the critical competencies of individuals may be judged, including scopes of examinations, equipment and materials
7. Identification of weighted criteria for each job or task along with preparation of rating scales and scoring procedures
8. Standardization of testing procedures
9. Pilot testing of the instruments
10. Analysis of data
11. Revision of tests as needed
12. Field testing of examinations
13. Analysis of test and demographic data
14. Preparation of norms, reliability measures
15. Preparation of a technical manual
16. Research reports and studies

17. Establishing support systems, facilities, staff, operations
18. Undertaking steps for test revision
19. New test development activities
20. Special studies, stability, applications to other populations
21. Major revisions through repetition of the process
22. New development through redesign
23. Comparative analysis of alternative forms
24. Data collection and analysis
25. Test revision
26. Reporting and implementing new developments

The closer one can duplicate reality in a performance test, the better the measure will be. The actual operation of machines, apparatus, instruments and tools used on the job should be included. The step-by-step procedures involving designing, cutting, forming, turning, shaping, and assembling units into components has to be demonstrated as well.

In situations where troubleshooting represents a major part of the occupation, such as in the electronics field, the step-by-step procedures for locating the malfunctions in equipment and instruments should be documented by the examinee. The student should also demonstrate his or her ability to remove and replace defective parts or components, as well as calibrating and maintaining instruments used in the occupation. To illustrate the approach, the machine tool trades will be discussed.

The machine trades occupations can be divided into divisions such as layout, benchwork, machine tools, heat treatment and so forth. Once these divisions have been made, it is necessary to identify the critical competencies required to perform tasks and jobs within each division successfully. This analysis will reveal that there are similar types of skills required to operate a different piece of equipment. It is this recognition which will help the test developer synthesize tasks and gain economy in terms of the number and types of jobs required to demonstrate mastery of a competency. Table 3 lists some of the skills within one division of the machine trades area.

Table 3

Selected Skills Necessary Within a Divisional Area

Major Division

Subdivisions

calipers (use and application)

steel rules

protractor

radius gauge

micrometers

hole gauge

vernier calipers

height gauge

dial indicators

layout and inspection

measurement of surface finish

blueprint reading

sketching, and making of technical drawings

use of layout fluid

layout of work piece

precision layout

surface plate

vernier height gauge

comparator

In the machine trades area, the critical competencies included under layout and inspection, could include jobs that require such functions as layout of work, including centers, reference, contour and dimension lines, surface preparation using common hand and measurement tools, surface plate and other holding or clamping devices, precision tools and gauges, testing and inspecting with precision inspection tools, precision blocks, gauges; indicators, hardness testers, and use of a comparator. Therefore, the performance job selected should sample the procedures that require a working skill using the measurement tools listed.

The specific tool or procedure selected would depend on the level of sophistication needed to be judged competent in a given job situation. The criteria for assessing skill proficiency should include both process and product measures and a rating scale used by an evaluator to observe the subject. Performance of an individual taking such an examination might include such criteria as:

Process (These criteria provide standards related to how each candidate undertakes to accomplish the job, the methods and techniques used.)

1. Handling of layout tools

2. Planning of layout procedure

3. Layout process

Product (These criteria provide standards related to what each candidate accomplishes, the outcomes.)

1. Accuracy

2. Precision

3. Time

Note: As the experience of candidates increases, product measures provide more important indicators of competency.

When the ranking of individuals is important, ratings of performance on each criterion should be noted. When absolute mastery is essential, a check list may suffice. The first approach allows for norm referencing while the second can be criterion referenced.

To the extent possible, various weights can be given to criteria. The weights should reflect the importance and frequency of those criteria in relation to the competency being examined. The more important aspects of the occupation should be weighted higher than less important competencies.

After the initial design of a performance test has been prepared, the test should be reviewed by knowledgeable people in the field. This content validity step will increase the probability that the content of the examination and the criteria are appropriate and adequate. In essence, this would be the second major validity check of the examination; the first being agreement among experts on the list of critical competencies.[2]

Standardization

When the consultants have agreed on the test, all necessary information and materials to conduct the examination should be prepared in a manner that will permit identical administration of the test. Items to consider include:

1. A handbook, providing directions for the examiner as well as for the student.
2. A set of jobs that will be required by each candidate including specific weighted criteria, and the amount of time usually required to complete each subunit of the test. (A subunit represents a job containing a series of specific job competencies)
3. A scale stipulating specific criteria.

Having assembled these materials, it is now desirable to pilot test the examination under a variety of conditions. For example, the test might be given to:

- People who are currently employed in the occupation
- Students completing training in the occupation
- Students starting their training in the occupation

The developer should observe if the items are indeed functioning properly. Are there real differences in the scores achieved by the different populations? Students who are beginning in an area should do significantly less well than those who have been in the job for some time. If these conditions do not prevail, then modifications to the test instrument are required.[3]

The individuals who will be used as evaluators should be given an opportunity to take the performance examination themselves. This type of hands-on experience will point out to the evaluators some of the problems likely to be encountered by examinees. All of the conditions required for the administration of the test should be the same for each administration of the test. Because performance tests usually require the use of local equipment or tools, some variance in scores cannot be completely controlled. Their effects can be reduced if candidates are checked out on the equipment before the test or if they are permitted to use their own tools.

Validity, Reliability, and Norms

In addition to the content validity and agreement of judges, the results obtained from performance examinations should be compared to other measures of student achievement, such as a student's grade point average. A high correlation, in this case, would provide a measure of the test's criterion validity. Supervisory ratings achieved by people in the workplace could also be compared with the student's performance. Basically, if these measures were taken at approximately the same time, they would be an indication of the test's "concurrent" validity. The integration of several factors to measure an abstract concept is called "construct" validity. Developmental efforts regarding the identification of such traits can be incorporated in performance tests, if desired. Validity of the test is the degree to which the test measures what it was designed to measure.[4] (See Table 4.)

Table 4

Types of Validity and Their Application

Type	*Application*	*Test*
1. Content	1. Test of skill and training	1. Samples desired domain, judges' consensus.
2. Criterion-related	2. Prediction of future, based on current data	2. Correlation between scores and criterion, measured over time.
3. Concurrent	3. Prediction at a specific time	3. Correlation between scores and criterion. Tests of other measures obtained at the same time.
4. Construct	4. Measurement of a scientific idea or factor; abstractions	4. Explanation of variance through experimental design.

Note: Performance tests are usually validated using content analysis. Other forms of validity also can be applied.

When evaluator judgments are required, such as in the case of most performance tests, a measure of interrater reliability is desirable. Administrative costs associated with obtaining such measures are high.

A recent study at NOCTI demonstrated the efficacy of the use of Cronbach's Alpha measure of reliability for determining the internal consistency of performance tests. There are other types of reliability that, likewise, apply. For example, measure of stability can be obtained by using the test-retest approach, or measures of equivalency can be obtained comparing alternate forms of the test by means of correlation coefficient. Reliability is a ratio of the true variance in a set of scores compared to the total variance obtained. True variance is what remains after all of the factors which may contribute to error are explained or controlled. Errors may result from many sources such as: ambiguity of materials, test administration, inconsistencies, examiner biases, and subject anxiety about test-taking to name a few. A summary of some of the approaches related to obtaining measures of reliability for performance tests is presented in Table 5.[5]

A performance test that is both valid and reliable requires an application to a sample of the population it purports to measure in order to establish norms. The field test can provide the data from which standard scores may be derived. The scores can be reported as percentiles or in some other appropriate form such as a "T" score, where the mean equals 50 and the standard deviation equals 10. In addition to overall performance test score norms, it may make sense to develop subscores for diagnostic reasons. Such measures can provide counselors and educators with a more precise indicator of a student's accomplishment as well as a measure of unmet need within a specific area of understanding.

The method of providing standards, which has been described, is called norm referencing. The standard scores are based on the distribution of scores of a sample for a specific population. The norms provided by test publishers usually pertain to a large area; when feasible, local norms should also be prepared.

Performance tests also can be scaled using other approaches. One such approach is criterion-referenced norming. In this case a specific level of mastery is required for success. The most widely recognized examinations that use this concept for norming are the tests given to people who want licenses to drive a motor vehicle. Because performance tests by definition must be content valid, they can be scaled in a criterion-referenced mode as well. This is because criterion-referenced tests also have to be content valid.

The Rasch method of scaling may also have opened other ways for performance test developers to scale tests. This scaling method is based on factors independent of population considerations. In essence, the technique provides data related to the percent of students on a specific level of development, who would be expected to respond correctly to the tasks. By testing students at different levels of achievement, one might arrive at a task characteristic curve which then could be used as the standard.[6] Conceivably, a single performance test could be scaled, applying the three methods in one instrument. The users would then select the norm that best fits the purpose for which the examination is being applied. The norm-referenced approach has become the acceptable standard for most tests. With time, especially as refinements occur regarding related test theory in terms of the criterion-referenced and Rasch models, the use of these newer techniques should find wider acceptance. Therefore, their application should become more common, especially in the area of performance testing.

In instances where the same performance test may be applied to different populations, it is appropriate to provide norms for each of the groups. Under ideal conditions, the developmental

Table 5

Some Approaches Towards Obtaining Measurements of Reliability for Performance Tests

Type	*Application*	*Test*
1. Test-Retest	1. Second administration of identical test where setting may have an effect.	1. Coefficients of stability.
2. Alternative forms	2. When there is a need for more than one test to measure the same performance.	2. Coefficients of equivalence.
3. Single form	3. When measures are to be obtained from one test.	3. Coefficients of internal consistency.

process for tests should include an analysis, describing the effects of the test among people of different sex, race, age, training, and experience. During the initial stages of test development, such analyses may not be possible. However, data of this nature should be obtained as the applications of the test become more diverse. In situations where equal employment opportunity laws apply, the data may be required. Furthermore, good practice requires that when subscored norms are provided, the information regarding the validity and reliability of the resources should also be included.

A caution needs to be raised regarding generalizations from too few criteria in a subset. On performance tests, work at NOCTI has revealed that at least four process and four product criteria are needed to obtain an acceptable level of internal consistency, a reliability of .90 or better. Generally, performance test makers have provided too few criteria or too many. Too few criteria may result in an inadequate measure of the test taker's true score. When too many criteria exist, the scales become difficult to administer which may result in increasing the rater's bias.

Cut Off Scores

There are no universally applicable methods for determining a cut off criterion. Frankly, it depends on many factors. It may be based on a probabilistic model. The cut off might be related to supply and demand for a given occupation. In situations where there is a large demand and a small supply, a more liberal criterion might be used; and the reverse might be considered under appropriate conditions. If a high degree of skill is required to demonstrate competency, then the cut off should reflect that level regardless of market conditions.

What is important in establishing a cut off is the rationale for determining and considering when a point must be clearly understood that it may be defended if necessary. Once such a cut off point has been established, the results of examinations should be monitored. This will ascertain whether or not the measures are providing weights for meaningful decision-making. It is only through constant reappraisal that appropriate cut off scores can be maintained.

Another concept to remember is that a test's cut off score must be fair; fair to the candidates taking the examination and to the people they will serve in the occupation. The measures obtained from a performance test represent an estimate of an individual's performance under a given set of conditions. They cannot represent all of the characteristics required to perform a given task adequately. However, if the performance test has been constructed using common practices, there will be a high probability that scores achieved on the examination will reflect the individual's ability to perform successfully on the job and in the domain examined.

Test-Related Materials

In addition to standardizing a test and obtaining measures of reliability and validity, it is important to provide data about the test to the users. This information may help the user make decisions about the appropriateness and adequacy of the examination as well as providing directions for test administration, scoring, and interpreting the results. A manual should be designed to convey pertinent information to users of the test.

Reference to studies that involved the use of the test should be included, such as studies concerned with measures of validity and reliability under varying conditions. Any claims made by

the publisher about the test should be substantiated in the information contained in the manual or by reference in the manual to a score that would support the claim. In addition, information regarding how the test was developed should also be included."

Since testing is a dynamic activity, these manuals should be revised and updated as research and conditions warrant. Information regarding how to interpret the exminations should be included along with warnings to the user regarding possible misuses of the information obtained through examinations. If the performance test norms were developed for use with a specific population, they may not be applicable to other populations. Information regarding specific applications and purposes of the examination should be explicitly stated. The manual should identify the qualifications needed in order to administer the performance test. The qualifications of the evaluators are as important as the test itself when it pertains to occupational competency assessment.

Directions for administration and scoring a performance test should be clear so that the examination can be similarly conducted in all settings. One problem in preparing examinations is that the laboratories or shops where tests are conducted are different. Under strict standardization process, it would be generally held that candidates taking the examination should be required to perform the test on the same piece of equipment. Although manufacturers tend to produce machines of comparable design, tests, out of necessity, will be conducted using different makes of the same tool. Therefore, in the instructions to the evaluator, a notice should be given that equipment having similar specifications to the suggested standard may be substituted. Skilled workers, with a minimum amount of instruction can function effectively on equipment manufactured by different companies. In situations where candidates may be unfamiliar with a specific piece of equipment, they should be given an opportunity, prior to the examination, to become familiar with the controls of the equipment. They should also be permitted to operate the equipment for a brief period of time before the start of the test.

When the examiners are required to score their own ratings, there should be procedures presented in the manual with enough detail to minimize the probability of scoring error. In situations where the scoring is to be accomplished by a test publisher, it is recommended that the evaluation rating sheet contain, in addition to the numerical assessment, some space for general statements or comments regarding the overall performance of each candidate by the evaluator. This information can be useful as an internal control. A candidate's total numerical score should be in agreement with general statements made by the evaluator. For example, if in the evaluator's numerical rating, the rating turns out to be extremely high, his general comments should be consistent with this measure. If this were not the case, a follow-up should be initiated to correct this apparent discrepancy.

Standardized measures of central tendency, standard deviations, standard errors, median, and validity and reliability and correlation coefficients with their standard errors of measurement should be contained in the manual with the fundamental data. Demographic data and sample size from which the data was derived should also be reported in the manual. All of the data reported in the manual should help the potential user determine the suitability of the test in terms of the particular application as well as to assist in the interpretation of scores.

Since "skill" is a relative term, each of the criteria selected could be judged on a rating scale containing at least three levels such as extremely competent, average and inept. Along with the basic information contained in the manual, it should be stated that local norms may vary from the norms that are published in the manual. When populations are large enough, it may be desirable to have local norms.

In addition to norms for individuals, class or program norms can be derived The use of performance tests within a school setting may require the derivation of special measures used for the purposes of analyzing group achievement as a measure of teacher effectiveness. Performance tests may be used to assess the performance of a group or program. This requires using the mean source of each class of students in a given program instead of individual student scores. When such norms are provided, the user should be informed of their derivation and application. Whatever approach is used, it is important to provide the user with the standard errors so that the precision of the measurement may be understood.

Test Revision

Once the test has been used, the test developer must continue, on a periodic basis, to update and improve the instrument. Although generally the critical competencies within any occupation do not change radically from year to year, important developments do appear that must be considered. The magnitude of these changes varies among occupations. For example, in the printing industry during the last twenty years, there have been tremendous changes in technology. The same holds true in the field of electronics. However, changes in fields such as carpentry or masonry tend to occur at a substantially slower rate of development. Therefore, the rate of change on any performance test measuring competencies in these occupations would not vary greatly from year to year.

What is the most appropriate time to change jobs on performance tests? Rather than be completely random regarding when to change some items, NOCTI, for example, deletes a competency when it is not being used in at least 25 percent of the field and adds new competencies after the practice has been adopted by at least 25 percent of the field. The 25 percent is an operational standard that can be modified up or down depending on circumstances. When a replacement job has been selected, if the time required to accomplish the task and its total value on the test is similar to the item being removed, the change may not seriously affect the cumulative norms. However, any change within an instrument must be examined to see whether or not the change could cause a change in the norms and thereby invalidate the standard. New norms are usually needed when jobs are changed. In cases where examinations have high reliability, small changes on the test do not appear to alter outcomes. It is therefore possible to update examinations and use the cumulative norms without necessarily being too concerned about problems of independence. However, if this practice exists, it is best to monitor test results to make certain that significant differences do not occur, since what may appear to be a small change could affect results in significant ways.

Performance tests may not cover the latest developments or all of the techniques employed by individuals engaged in a specific occupation. However, generalizations about a person's skill can still be valid. Just because someone has knowledge does not necessarily indicate competency. For example, a student may know all of the latest techniques, and yet some of these techniques may still have to gain acceptance in the field. The reverse may also be true; the field may be well ahead of the training institutions. Therefore, the performance measurement does not reflect competency unless the examination is based on the current practice in the field The testing should relate as directly as possible to reality. This direct parallel with the world of work provides specific information regarding student accomplishments in terms of the needs of employers.

Since the tasks are based on reality, performance tests can be used to evaluate the quality of programs as well as to point out areas in need of improvement.

There will be situations that preclude using a direct experience. For example, in the case of flying an airplane, a simulator may be the best way to test initially for the skill level rather than to permit the student to directly control the flight of a plane.

When feasible, alternative performance tests should be provided. Several versions of the same instrument is a help in regard to test security. Although it may be highly desirable to have other exams, the cost involved in such development is high, and the difficulty of arriving at equivalent forms sometimes precludes their application.

Performance tests appear to provide measures of achievement that are not biased due to race or sex. Because of the importance test scores have on the future of an individual and society, concern is often raised about test bias due to race or sex. NOCTI has found that scores derived from performance tests tend not to contain these forms of error variance. Test results should be communicated clearly. This suggests the describing of the confidence interval around a test score, rather than just reporting the point estimate of the measure. A report of scores should be accompanied with all the necessary information required to interpret the measure.

A Few Concluding Comments

Vocational instructors have always used performance tests. The basic difference between their approach and the one described in this chapter is that the test development procedures followed by instructors normally result in larger error terms. Standardized tests are more likely to reduce the size of the error in the student's score.[9] Therefore, they provide a better estimate of student's true achievement level along with obtaining comparable measures across programs.

The performance test samples an individual's ability to perform jobs and tasks that are judged to be critical and important within a given occupation. They may take the form of real work or a simulation of work. Regardless of what form they may take, they should be as realistic as possible. Performance tests provide a way to assess psychomotor skills as well as to provide for an alternative way of examining a person's problem-solving ability. When coupled with other measures of achievement, they provide valuable insights regarding an individual's ability and a program's effectiveness.

Performance tests are simply another method of measuring skill attainment. These tests, themselves, must meet general standards for educational and psychological testing.[10] In the past, this has not been effectively accomplished. With advances in test construction and measurement theory, it is now feasible to create effective and efficient performance instruments. Because techniques are now available to standardize such tests, their usefulness will continue to be appreciated, and their application will continue to expand. The age of standardized performance tests has only just begun, and its future looks promising.

Notes

[1]Panitz, A. and Olivo C. T., *Handbook for Developing and Administering Occupational Competency Tests*. New Brunswick, N.J. Rutgers University, 1971

[2]Lennon, R. T. "Assumptions Underlying the Use of Content Validity" *Educational and Psychological Measurement* (1956): 294-304.

[3]Anastasi, A. *Psychological Testing*. 4d, New York: MacMillian, 1976.

[4]Cronbach, L. J., and Gleser, G. C. *Psychological Tests and Personnel Decisions*. 2d. Urbana, Il., University of Illinois Press, 1965.

[5]Klein, R. S. and Pfeiffer, S. "Measuring the Internal Consistency of Selected NOCTI Performance Examinations," Paper, Albany, N.Y.: National Occupational Competency Testing Institute, 1979.

[6]Forster, F. "The Rasch Item Characteristic Curve and Actual Item Performance." Paper, San Francisco: American Educational Research Association, April 1976.

[7]United States Equal Employment Commission, et. al., "Uniform Guidelines on Employment Selection Procedures." Federal Register, (1978): 38295-38309.

[8]Davis, F. et. al. *Standards for Educational Psychological Tests*. Washington, D.C.: American Psychological Association, 1974.

[9]Thorndike, R. L., *Educational Measurement*. 2d, Washington, D.C.: American Council on Education, 1971.

[10]Tunkel, L. S. "Occupational Competency for Quality Vocational Education". Monograph, Columbus, OH.: State Education Department, 1979.

Comments on the Technical Issues in Performance Testing

Samuel A. Livingston
Center for Occupational and Professional Assessment
Educational Testing Service
Princeton, New Jersey

These two papers raise a number of technical issues in the development and use of performance tests:

- How should the test maker select tasks for a performance test?
- Should a performance test evaluate the student's procedure or only the product of the task?
- How can the student's performance be translated into a test score?
- How can we reduce the influence of irrelevant factors on the student's score?
- What types of reliability are particularly important in performance testing?
- What types of validity are particularly important in performance testing?
- How should we set the pass/fail cutoff on a performance test?

Klein suggests that the main consideration in selecting tasks for a performance test is that the tasks should adequately sample all the skills that are to be tested. This suggestion is good as far as it goes; redundancy in testing is often a luxury that performance testers cannot afford. But what should the test maker do if there is not enough testing time to test all the skills? I would suggest that there are two considerations: (1) the consequences of allowing someone to remain deficient in a particular skill, and (2) the extent to which the skill can be tested by other, less time-consuming and less costly methods.

Klein suggests that both process and product should be evaluated in a performance test. This advice is usually sound. Concentrating entirely on the product and ignoring the process can be dangerous, especially when safety precautions are involved. Process evaluation is also important when a bad procedure yields a bad product only part of the time. But if no safety precautions are involved and if wrong procedures always show up in the product, an evaluation of the product may be sufficient. In other performance tests, it may make sense to base the evaluation entirely on the process. The product of the task may be difficult or impossible to observe. The quality of the product may depend heavily on factors that cannot be standardized. Or the product may be a joint effort of two or more persons, only one of whom is being tested. In these cases, an evaluation based entirely on the process is quite appropriate. But in many performance tests, it makes sense to evaluate both process and product.

Both papers deal with the problem of converting performance into a test score. Perloff suggests several weaknesses of rating scales but does not offer any alternative. Klein suggests that performance testers use rating scales for ranking students, using check lists only "when absolute mastery is essential." Actually, a highly detailed check list may provide enough information for ranking students, as well as for determining their mastery in an absolute sense. Also, a check list requires less judgment on the part of the observers and thus reduces the extent to which the student's score depends on the observer's individual standards (and the observer's mood at the time the of the test). The completed check lists also provide a detailed, descriptive record of students' performance, for diagnosing student's (and instructors') weaknesses, and for documentation in case of a disputed score.

Both papers offer several specific suggestions for reducing the influence of irrelevant factors. In brief:

- Standardize the testing conditions
- Give the observers detailed instructions.
- Train the observers
- Use more than one observer if possible
- Have the observers record their observations as soon as possible after making them.

One additional technique that is often helpful is to give the observers examples of adequate and inadequate performance for each aspect of the task requiring the observer to make a judgment. These examples should illustrate borderline cases if possible. That is, the example of inadequate performance should be nearly adequate, while the example of adequate performance should be just barely adequate. Examples of this type provide a clear standard for the observers to use in judging the students' performance.

Reliability is the level of agreement between test scores that would be the same if the scores were free of the influence of irrelevant factors. In performance testing, the most important of these irrelevant factors is usually the selection of a particular observer. Therefore, the most important type of reliability is inter-observer reliability. To determine the inter-observer reliability of a performance test, you must try it out with at least two observers observing the same performance. If the test involves an evaluation of the student's procedure, both observers will have to observe the student at the same time (unless the performance is recorded in some way, e.g., video-tape). Other types of reliability may also be worth investigating, e.g., short-term stability, or alternate-forms reliability (where the alternate forms contain different tasks selected to test the same skills).

Internal-consistency reliability statistics such as "KR-20" or "alpha" are often irrelevant to performance tests. They should not be applied to the checkpoints on a check list, because the checkpoints are not a sample from a much larger universe of possible checkpoints. They are not interpreted in terms of some underlying trait. They represent only what are they are—the most important observable aspects of the task. However, there is one case in which internal consistency reliability statistics would be relevant to a performance test. This is the case of a performance test that contains several tasks, all intended to measure the same general abilities. In this case, the "items" would be the tasks, not the individual checkpoints.

Validity is the extent to which a test does the job it is being used for. More than any other kind of test, a performance test is a direct measure of the skills it is intended to test. Therefore, the most relevant type of validity is content validity. Even the harshest critics of content validity concede that it is relevant when the information resulting from the test is expressed in "the strict behavioral language of task performance."[1] However, if we intend to draw inferences about the student's performance in situations unlike those included on the test, criterion-related validity may also be relevant. (The concept of "consistency validity" introduced by Perloff does not seem very different from content validity.)

Perloff is correct in asserting that validity is the most important characteristic of any test. A test that does not yield valid scores is worthless as a measuring device. However, there is often a trade-off between validity and efficiency. It may be necessary to sacrifice some degree of validity to achieve a gain in efficiency, which is what we do whenever we use any form of simulation in a performance test. Often the most difficult decisions in developing a performance test involve the trade-off between validity and efficiency. The real world forces us to do our testing with limited resources (time, money, and so forth) and without risking the safety of the students or other persons. Validity is the main thing, but it is not the only thing.

More than any other type of testing, performance testing offers an opportunity to choose cutoff scores in a way that most experts would acknowledge as correct, or even "optimal". Any method of choosing a cutoff score involves judgment. What is important is that the judgments must be made in a way that assures their meaningfulness and that they must be made by persons who are qualified to make them. Probably the most meaningful type of judgment is the direct judgment of examples of performance as acceptable or unacceptable. In most other kinds of testing, it is difficult to get meaningful overall judgments of students' performance; in performance testing it is easy. Judges' standards will vary, but these differences will tend to "average out" if several different judges participate in the process. By analyzing the students's test scores together with the judgment of their performance, we can estimate the probability that a student with a given test score would be judged (by a randomly selected judge) to have performed acceptably.

To use these probability estimates to set a cutoff score, we (i.e., somebody) must make one other type of judgment. There are two types of decision errors we can make. We can pass a student who deserves to fail, and we can fail an student who deserves to pass. What is the relative seriousness of these two types of errors? We will never be able to eliminate decision errors completely, as long as there is any test score at which some students are judged acceptable and others unacceptable. The best we can hope to do is to minimize the total harm from the errors we will make. When we know the probability of each type of error at any given test score level, and the relative seriousness of the two types of errors, we can choose a cutoff score that is "optimal" in this sense.

Notes

[1]S.J. Messick. "The Standard Problem: Meaning and Values in Measurement and Evaluation," *American Psychologist* 30(1975): p. 955-66.

CHAPTER FOUR

LEGAL ISSUES

Questions frequently arise in the field of education that often find themselves to be part of broad legal issues affecting the delivery of all types of educational services. The institutionalization of performance testing in vocational education programs, for example, brings with it a series of legal concerns to which teachers and administrators must be sensitive. Paul L. Tractenberg's paper opens Chapter Four with an overview of the legal implications of performance testing. He begins by identifying the major legal provisions—due process, equal protection clauses, state education clauses, federal and state education statutes, federal and state regulations—which may prove relevant to performance testing. Tractenberg then applies the legal theories to a series of keynotes on minimum competency testing that have been adapted to performance testing in vocational education.

The second paper in this chapter, by Diana C. Pullin, identifies lessons to be learned from the minimum competency-testing movement. She discusses the question of accountability through performance testing from a legal perspective and then focuses on several legal areas which should be of concern to vocational education. The question of fundamental fairness is raised, as is the fundamental flaw in one minimum competency testing program, and some recommendations for fundamental fairness in vocatioal education performance testing programs are identified. The paper then discusses the potential for unlawful discrimination and the right to privacy. Finally, recommendations are offered to the reader

William G. Buss provides another look at the legal issues from a third perspective in the Comments paper. He emphasizes "some of the legal ambiguity and related interaction between law and education that is involved in the material considered in these papers."

Legal Implications of Performance Testing in Vocational Education: An Overview

Paul L. Tractenberg
Rutgers School of Law
Newark, New Jersey

During the past several years, the minimum competency testing movement has swept across the country. It has left controversy in its wake. Proponents laud its potential as a vehicle for increased educational accountability;[1] critics attack its basic premises and the feasibility of implementing it effectively.[2] Some observers believe that the movement has already peaked, and that the educational reform pendulum will begin to swing in the opposite direction.[3] For the moment, though, some form of minimum competency testing program is in effect in about three-fourths of the states.[4] The implications of these programs for school systems, for education professionals, and for students are potentially great.

One arena in which those implications are being explored is the courts. Students and parents have sought judicial intervention to prevent injuries that they allege will result from particular minimum competency testing programs. The first important decision—the Florida federal district court's decision *Debra P. v. Turlington*[5]—has been handed down. Several other significant cases are pending[6] and more are certain to be filed. The impact of these cases on the present and future status of minimum competency testing is likely to be substantial.

Judicial involvement in matters of pupil assessment is not new.[7] To a considerable degree, the courts have sought to defer to the educational authorities where the issues raised by the cases were whether the assessment instruments were appropriately developed or administered,[8] or their results were appropriately used.[9] But, in some cases, the constitutional rights of students were so clearly and substantially implicated, or the actions of the educational authorities were so deficient, that the courts saw no alternative but to intervene.[10]

It is against this backdrop that the use of performance testing in vocational education must be considered. Performance testing in vocational education has significant parallels to minimum competency testing in general education. Indeed, the momentum generated by the latter undoubtedly has contributed to increased interest in the former; peaking of the minimum competency testing movement, or adverse court decisions, therefore, would have implications for performance testing. But performance testing in vocational education has a history and relevance which are independent of the minimum competency testing movement.

This paper has three purposes, each the subject of a separate section: (1) to provide a brief overview of legal principles and provisions that are likely to be relevant to performance testing in vocational education; (2) to describe the major policy decisions involved in developing a performance testing program and to assess the legal implications of each; and (3) to predict legal developments and consequent policy directions. The work of Brickell in articulating the seven keynotes of competency testing[11] and of Ahmann in applying them to performance testing in vocational education[12] provide a convenient organizing framework.

An Overview of Relevant Legal Provisions

There are seven categories of legal provisions that may prove relevant to performance testing developments. Four are constitutional in origin: federal and state due process clauses, federal and state equal protection clauses; federal and state clauses protecting privacy and freedom of belief; and state education clauses. The fifth is statutory—those provisions of federal and state statutory law that directly or indirectly bear upon the establishment and operation of performance testing programs in vocational education. The sixth is regulatory—relevant policies, rules, and regulations of the federal and state education authorities. The seventh is the "common law," legal principles evolved through the litigation process. Each of those sources of law will be considered briefly.

1. *Federal and state due process clauses.* The Fourteenth Amendment to the Federal Constitution and most state constitutions contain a due process clause. The federal clause provides that no state[13] shall "deprive any person of life, liberty or property, without due process of law." The judiciary has construed due process to have substantive and procedural aspects.

Substantive due process, still in existence although significantly diminished in legal importance,[14] requires that the action of the state be rational and reasonably related to a legitimate state objective. If, for example, it could be proven that performance testing was evaluating students on materials or skills never taught in the vocational program, students who failed on that test to demonstrate their proficiency might credibly assert a violation of their right to substantive due process.[15]

Procedural due process requires that the state act in a fair manner when it deprives a citizen of liberty or property. In connection with a performance testing program, procedural due process might require, for example, a procedure under which students with "failing" scores be permitted to challenge the scoring of the test, the qualifications of the test administrators, or the validity of the test itself. It might also require adequate phase in time for a performance testing program that imposed substantial sanctions. The absence of adequate phase in time was one of the bases for the *Debra P.* Court's four-year deferral of the diploma sanctions under the Florida minimum competency testing program.

Both substantive and procedural due process require a showing that a person has been deprived of liberty or property by action of the state. Students could assert that denial of a diploma, or of promotion or graduation, or of full access to the job market, as a result of performance testing constitutes a deprivation of "property." Courts have found that students have a property interest in their education such that physical exclusion from school, even for a short time, requires due process procedures.[16] In the *Debra P.* case, the court found that students would be deprived of a property interest by a minimum competency testing program that determined whether they would be graduated.

The *Debra P.* court also found that the minimum competency testing program deprived "failing" students of a liberty interest by stigmatizing them as incompetent or ineligible for promotion, graduation, or a regular diploma.[17]

It should be remembered, though, that proof of deprivation of a liberty or property interest in itself does not condemn the state's action; it obligates the state to act fairly and rationally. Indeed, during the past several years there has been something of a trend in the federal courts to expand governmental prerogatives and discretion, and to afford correspondingly reduced judicial

protection to aggrieved citizens.[18] At least some state courts have resisted this trend in interpreting their state constitutional due process clauses.[19]

2. *Federal and state equal protection clauses* Equal protection is a constitutional principle related to due process. Both require governmental rationality and fairness in treatment of citizens. The federal equal protection clause also derives from the Fourteenth Amendment. It prohibits the state from denying "to any person within its jurisdiction the equal protection of the laws."

The equal protection clause tends to focus on state action with respect to groups rather than individuals. A challenger of state action must show that it classifies persons and provides differential treatment to them without adequate justification. In the federal courts, as well as in some state courts, the burden of justification required of the state for differential treatment increases with the importance of the interest subjected to such treatment. The courts speak of "fundamental" interests imposing upon the state the burden of showing a compelling reason, for, and no available alternative to, the differential treatment. This "strict scrutiny" approach is also invoked by "suspect" classifications, such as those based upon race. Interests of lesser importance or classifications not based on a suspect characteristic result in a lesser burden on the state—perhaps only the need to prove that the classification is rational, even if not the best means to achieve the state's objective. In recent years, an intermediate approach has been developed to deal with certain kinds of cases, and a "sliding scale" approach, in which the importance of the citizens' interest is balanced against the significance of the state's justification, has been advocated.

An equal protection challenge to performance testing in vocational education likely would proceed along one or both of the following lines: (1) that, to the extent black or Hispanic students were disproportionately represented among those failing to demonstrate proficiency, the program classified students racially or ethnically—a suspect classification—and should be subjected to strict scrutiny; or (2) that the program lacked even a rational basis because, for example, the test was invalid[20] or covered material or skills not taught in the schools.[21] The argument that strict scrutiny should be applied because of the fundamental nature of education is unlikely to succeed in the federal courts. The United States Supreme Court ruled to the contrary in 1973.[22] Several state courts have reached a contrary conclusion, however, under state equal protection clauses.[23]

Recent U.S. Supreme Court decisions also have created problems for an equal protection challenge based upon racial or ethnic discrimination. The Court has ruled that a statistically disproportionate effect, while relevant, is insufficient to demonstrate a racial or ethnic classification.[24] Challengers of state action must prove, by direct or circumstantial evidence, that there was an intention to create such a classification. That may be a formidable task in the context of a performance testing program. On the other hand, if the particular state or school system previously has engaged in unlawful discrimination, it may have an ongoing duty to eliminate the effects of that prior discrimination. In that situation, even a neutral classifying device could be found deficient.[25]

3. *Federal and state freedom of belief and privacy provisions.* The scope and content of some performance tests may raise significant issues under the First Amendment's right to freedom of expression and belief, and the Fourteenth Amendment's implicit right to privacy, and their state constitutional counterparts. These problems will arise primarily from the inclusion in performance tests of items that assume or inquire into values, attitudes, or characteristics considered relevant to job success, such as punctuality, respect for authority, and ability to get

along with co-workers. There are two areas of concern: (1) in order to demonstrate proficiency, the student in effect must subscribe to certain values; and (2) the student may be required to reveal confidential personal matters.

There are no judicial decisions which provide definitive guidance about how these issues will be resolved in a challenge to a performance testing program. An important line of Supreme Court decisions* does afford students with some protection against the efforts of school authorities to have them believe in a certain way or to express certain beliefs. In performance testing, however, student values, attitudes or characteristics may be highly relevant to predicting success on the job. To the extent that such a prediction is an important element of performance testing, eliminating it might limit the validity of the testing. Thus a court will have to balance the respective interests carefully.

Another significant issue relates to the confidentiality with which performance test results are treated. If the results are kept confidential, the intrusion into a student's privacy is minimized somewhat. But an important purpose of performance testing is to provide prospective employers with information about the abilities of applicants. The invasion of privacy problems may be minimized if the students have to approve the dissemination of performance testing results. Ultimtely, however, the court may have to confront the question of whether there are limits to the state's power to inquire about a student's personal views and beliefs. It will do so by balancing the invasion of privacy occasioned by the testing program against the state's purpose in implementing the program.

4. *State education clauses.* Every state now has in its constitution a commitment to provide school-age residents with a free public education. About three-quarters of the clauses describe, to some extent, the required education.[27] These clauses may be relevant to, or the basis of, a variety of performance testing challenges.[28] For example, the absence of a performance testing program might provoke a challenge based on the state education clause. The argument could proceed as follows in a state with a "thorough and efficient" clause: The clause obligates the state to provide an educational program designed to equip students to function as citizens and as competitors in the labor market;[29] proficiency in vocational skills is essential for those purposes; establishment of a performance testing program is necessary to ensure that all students have an adequate opportunity to achieve such proficiency.[30]

State education clauses may also support challenges to particular performance testing programs. For instance, the levels at which proficiency standards were set could be challenged on the ground that they were not consistent with the state's obligation to provide a "high quality" or "thorough and efficient" education, especially if those education clause requirements had been construed to relate to the students' capacity actually to function in the postsecondary school work world. Challengers might argue that the standards were too low; performance at those levels would not permit students in fact to function adequately in employment.[31]

Another type of education clause challenge could be brought against a performance testing program that required or permitted different standards to be established by different vocational schools. Some education clauses expressly mandate a "general and uniform" system of education for the state;[32] others have been interpreted to require uniformity across district lines. Arguably, such clauses would be offended by a performance testing program that permitted a student's graduation or diploma to depend upon the district of residence or the school attended. On the other hand, educational home rule is a well-entrenched tradition in many states, including, paradoxically, some with uniformity clauses.

Finally, a state education clause challenge might be directed at the inadequacies of remedial programs for students who fail to demonstrate their proficiency. If a state has defined its educational mission to include pupil proficiency in vocational skills, then it must take reasonable steps to carry out that mission. Effective remedial education, once student deficiencies have been identified, is an important element.

5. *Federal and state education statutes.* Although at the present time there is no legislative parallel involving performance testing in vocational education to the minimum competency testing movement, statutes may be enacted which specifically provide for performance testing. In that event, the requirements of those statutes may provide legal bases for challenging particular performance testing programs. A possible line of legal argument is that the program, as implemented, does not comport with the statutory requirements. Alleged noncompliance may take many forms, ranging from blatant failure to meet specific requirements (e.g., failing to institute testing by a date specified in the statute) to more complex issues of qualitatively inadequate programmatic efforts (e.g., failing to provide educationally sufficient remedial programs for students who fall below the proficiency standards). Several legal challenges to minimum competency testing programs have raised these sorts of issues. For example, in one case, the challenge is based upon the schools system's alleged failure to comply with a specific statutory requirement to obtain parent, teacher, and student participation in the formulation of the program.[33]

Other more general provisions of federal and state education laws may be relevant, too. For example, there are statutes that provide guidelines for the operation of vocational programs,[34] that bar racial or ethnic discriminaton in education,[35] that provide for certain access to pupil records,[36] that assure citizen participation in educational policy making and governance,[37] and that regulate the education of special groups of students.[38]

Statutory challenges to performance testing efforts are likely to be narrower and focused on more specific aspects than constitutional challenges. By asserting a specific legislative standard, they will tend to reduce the court's concern about whether it may be substituting its judgment for that of another branch of government.

6. *Federal and state regulations.* Under many of the statutes referred to above, the responsible administrative agency has promulgated formal regulations or has issued interpretative guideines. In some states education regulations formally promulgated by state education authorities have the force of law. They can form a direct basis for legal challenges relating to performance testing programs in much the same way as statutes. Indeed, because regulations tend to deal with educational programs in greater detail than do statutes, they may provide a stronger basis for legal action. The more specific and detailed the prescription by a legislature or state education body, the more limited and mechanical the judicial intervention can be.

If, for example, state regulations provide in detail for a performance testing program pursuant to the authority of a more general statute, failure of the state or of the local vocational agency to implement that program fully can be challenged. The educational authorities may defend by asserting that despite the specificity of the regulations they should be permitted some flexibility, or they may seek to modify the regulations, or they may argue that the challengers have to exhaust available administrative remedies. All of these, however, are matters well within the traditional competency of courts to resolve.

In states where administrative regulations are not given the force of law, or in the case of administrative action, such as guidelines or policy statements, not having the status of formal

regulations, the substance of the adminstrative judgement should still have weight in legal preceedings. It represents the expert view of the state's educational authorities. As such, a court likely would find it highly relevant to an interpretation of broad constitutonal or statutory provisions.

7. *The "common law"*. The final source of law that may be influential in judicial consideration of a performance testing program is the "common law." Under the Anglo-American legal system this is judge-made law. Courts will tend to follow prior judicial decisions in similar cases under the doctrine of stare decisis. In confronting a new case, therefore, a court will consider, along with relevant constitutonal, statutory and regulatory provisions, the judicial precedent, especially cases decided in the same jurisdiction.

Many bodies of precedent are relevant to performance testing programs in vocational education. For example, as indicated previously, federal and state courts have dealt extensively with, and given content to, constitutional concepts such as due process rights of students, equal protection aspects of pupil classification by testing, educational segregation, and equality of educational opportunity. In many states related education statutes and regulations have been judicially construed. Beyond those possibilities, the courts have established certain legal rights independent of constitutional, statutory or regulatory provisions. Thus, students are entitled to be tested in a careful and appropriate manner by those who owe them a duty of care. School authorities which have failed to do so may be held liable for their negligence.[39]

Applying the Legal Theories to Performance Testing

Performance testing programs in vocational education may evolve in various ways. The differences in approach may be based upon differing perceptions as to what are the best public and social policies, educational program, administrative structure, use of available resources, and relationships to the job market. The purpose of this paper is to urge that legal considerations also should play a significant role in the development of performance testing. As a point of departure, I wil use Brickell's seven keynotes as Ahmann has adapted them to performance testing in vocational education. Ahmann also has added the "who" question at each stage in the developmental process. Thus, the keynotes become:

1. The skills and characteristics to be tested
2. The means of measuring them
3. The point(s) at which they will be measured
4. The number of proficiency standards which will be set
5. The level(s) at which these standards will be set
6. Whether the standards will be for school programs or students
7. The consequences of failing to achieve the standards
8. For all of the above, who will make the decision.

The skills and characteristics to be tested. A number of interrelated questions are raised by this keynote. They include the following: Are the skills and characteristics derived from the substance of the vocational education subjects? Are they derived from specific jobs (through job analyses) to which the vocational education subjects are related? Are they derived from categories of job? Are they derived from a broader idea of professional preparation, including Ahmann's concepts of "occupational knowledge" and job-seeking skills?[40] Are all the relevant skills and characteristics measured or a sample of them? Are values and attitudes to be included? Are general competencies to be included in the performance test or are vocational students required to take the separate minimum competency test used in the general educational program? Who determines the skills and characteristics to be tested (e.g., educators, employers or unions, students, parents or other citizens, or some combination of these)?

Consideration of the legal implications of the various alternatives may influence the policy decisions. In general, the most relevant legal theories are the substantive due process concept of rationality, the equal protection concept of nondiscrimination, the freedom of belief and privacy concepts, and the state constitutional, statutory, and regulatory requirements of a certain quality or quantum of education.

On one level, focusing on skills derived directly from vocational courses may comport easily with due process and equal protection concepts as long as: (1) the performance testing relates to subject matter that the students actually have had a reasonable opportunity to master; and (2) the selection of subjects taught or chosen for the performance testing is nondiscriminatory (in the sense that it is not skewed in favor of particular socio-economic, racial, or ethnic groups).

However, focusing on skills derived directly from vocational courses may pose greater legal difficulties under other concepts. State educational quality requirements, as well perhaps as substantive due process, may dictate that proficiency be defined in terms of skills actually required in the marketplace. In theory, vocational courses, more than any other school subjects, should be related to the marketplace. But that may not always be the case.

If the skills upon which the performance testing is based appear to be reasonably related to the job market in some sense, it is unlikely that a court will intervene because the skills are derived from categories of jobs rather than individual jobs, or from a broader idea of professional preparedness, or represent a sampling of relevant skills rather than all the skills involved. These are judgments about which the judiciary will tend to defer to the education officials, assuming that there is credible evidence that the task has been approached responsibly.

The courts are more likely to consider intervention if the performance testing gives substantial weight to personal values, attitudes, and other characteristics in addition to, or instead of, job-related skills. The risk of subjectivity and, ultimately, bias may be heightened by such an approach. Moreover, issues involving freedom of belief and privacy may be raised. Justifying the inclusion of such elements, therefore, is likely to be more complicated. On the other hand, if the educational authorities can demonstrate empirically that certain personal characteristics are closely related to successful performance on the job, they may be able to argue that the predictive validity of the performance testing is linked to inclusion of such elements. The courts will have to balance any infringement upon students' interests against the weightiness of the state's purpose.

The relationship between performance testing in vocational education and minimum competency testing raises further legal issues under the state's educational quality provisions. Generally, courts that have construed the state's obligation under such provisions have concluded that students have a right to an educational opportunity designed to equip them for

effective citizenship as well as for competition in the marketplace.[41] Mastery of basic academic skills may be relevant to both. Therefore, students in vocational programs will have to be included in the general minimum competency testing program, as they are in most states.

Finally, who determines the skills and characteristics to be tested may have legal implications. Certainly, the requirements of rationality under due process and equal protection concepts must be satisfied. Statutes or regulations might specify the procedures to be used in creating the performance tests, and their mandates would have to be met. Moreover, if the decisions actually were made by persons or agencies not officially part of the governmental structure, issues of improper delegation of authority would be raised.

The means of measurement. Brickell suggested four broad choices for measurement of student competencies that are applicable to performance testing in vocational education: (1) actual performance in job situations; (2) simulated performance in situations resembling the job; (3) performance in school programs; and (4) performance on paper-and-pencil tests.

The touchstone for evaluating these alternatives is the concept of validity.[42] Under both due process and equal protection doctrine, tests, of whatever type, must satisfy standards of objectivity, reliability, and validity.[43] Due process is implicated if the use to be made of the test threatens to deprive students of their rights to liberty or property. Evidence that the use of the test stigmatizes students who fail to demonstrate their competence or requires their attendance at remedial programs will be germane to an alleged deprivation of their liberty interest.[44] Denial of promotion or graduation based on the test results is the clearest support for deprivation of a property interest.[45] Even if a court could be persuaded that some students had been deprived of their liberty or property rights, the students still would have to prove that the test or related procedures were not procedurally or substantively fair.

An equal protection challenge would proceed most forcefully if a suspect classification were evident. At one point, a test's racially disproportionate effect—a far higher percentage of black than white students falling below proficiency levels—established a prima facie case of racial discrimination sufficient to shift a heavy burden of justification to the education authorities. Several years ago, however, the United States Supreme Court determined that an intent to discriminate, rather than merely discriminatory effect, had to be proven in order to establish a racial classification.[46] An intent to discriminate can be proven by circumstantial evidence, including statistical data, as well as by direct evidence.[47] It is still not clear, however, how heavy a burden that will place upon challengers of a performance testing program.

If, despite racially disproportionate consequences, no suspect classification can be established, and if the federal courts adhere to the view that education is not a fundamental interest, then the classification of vocational students into those who have achieved proficiency and those who have not can be justified by showing that it has a "rational basis." The validity of the testing instrument will still be part of the showing of rationality but the overall burden on the school authorities will be substantially lighter than under a stricter scrutiny approach. That is especially true given the recent tendency of the federal courts to defer increasingly to public officials' judgments.[48]

However, even if the performance testing is found to be racially neutral it may still be invalidated if the state or local educational system previously was found to discriminate against students and the effect of the testing is to perpetuate the effects of past discrimination. The federal district court in *Debra P.* found this to be the case with the Florida minimum competency testing program. Instead of invalidating the program, though, the Court merely deferred effectiveness of the diploma sanction.

The *Debra P.* court also dealt with many claims of test invalidity. In one of the weaker portions of the opinion, it concluded that although the test was flawed in many respects the inadequacies did not rise to the level of "constitutional infirmities."[49] This may suggest that if educational authorities can present evidence that they have attempted to deal with test validity concerns their efforts will not be struck down because they have fallen somewhat short of the "state of the art."

Applying these legal principles to the broad choices outlined by Bricker indicates that, in a general sense, paper-and-pencil tests may be more easily defended than the other alternatives. Although it may be more difficult to establish their predictive or face validity, the courts have not usually required such validity. Paper-and-pencil tests may be easier to validate in content or construct terms, and this is the direction of the court's primary focus. Moreover, paper-and-pencil tests may minimize the more obvious problems relating to objectivity and reliability that could plague tests based on actual or simulated performance.[50] Ahmann has described the difficulties, in terms of resources and personnel capability, that would have to be surmounted to develop and administer effective tests of actual or simulated performance. The courts will have little difficulty striking down a jerry-built performance test. This is not to suggest that, being the avenue of least legal resistance, paper-and-pencil tests should automatically be adopted. It does, however, reflect one of the realities that must enter into the decision.

The points for measurement. The purpose or purposes of the performance testing will determine, to a substantial degree, when the testing is carried out. The testing may serve a screening function for entry into a particular vocational program.[51] In that event, of course, the test would be given prior to entry into the program. If, on the other hand, the performance testing serves a certification function, it may be administered at or near the end of the vocational program. Finally, if the purpose is diagnostic and remedial for individual students, programs, or both, the testing will be administered periodically during the course of the program.

These purposes are not mutually exclusive. The choice of testing purpose and the related decision about points for measurement will be influenced by legal considerations. If entry into a vocational program is at issue, and the screening will disproportionately affect particular groups of students, equal protection questions will be raised. Due process questions may also be raised about whether the performance testing is an arbitrary means of screening individual students.

Central to both sets of questions are the validity of the particular performance test, discussed in the prior section; and the intention of the responsible education officials. In the latter connection, vocational educational professionals may have to deal with the argument that they attempt to limit entry into their programs to students who will be easiest to place in jobs. Critics have asserted that this had led to discrimination against black, non-English speaking and handicapped students.[52]

Similar legal issues will be raised if the purpose of the performance testing is certification. The sanction there may be withholding of promotion, graduation or a "regular" diploma, or identification of students as "lacking proficiency." The effect, in any case, may be ineligibility for, or reduced access to, future educational or employment opportunities.

Because of the weightiness of these consequences, the vocational education authorities' justification is likely to be subjected to careful scrutiny. This will include attention to the timing of the measurement. There should be adequate notice of the performance expectations and sufficient time and opportunity for students to meet them. A court that considered these matters also probably would require testing early enough to permit remedial efforts for students found to lack the necessary performance skills.[53]

If performance testing were for diagnostic and remedial purposes only, the burden of justification would be lightest. When and how frequently the testing was administered would be left to the discretion of the education officials, unless that discretion was exercised in a manifestly arbitrary or irrational way and some tangible harm to students could be proven. In this connection, the state education clause's quality standard might become relevant. Students might argue that the harm to them was that the program as structured could not provide them with adequate diagnostic and remedial efforts.

The number of proficiency standards set. There is a considerable range of policy possiblities concerning this matter. There could be a single statewide standard for all students in a particular type of vocational program, or there could be a separate standard for each student based upon perceived abilities, background and educational objectives. Between those poles are other possibilities—multiple statewide standards categorizing students by one or more of a number of possible criteria (i.e., demonstrated or projected intelligence, facility with English, existence of a handicap, socioeconomic background, nature of the particular school or school district and the community and job market that it serves, and the educational expenditure level); either single or multiple standards established region by region, district by district, or school by school for students within those respective jurisdictions; a combination of one or more statewide standards augmented by additional and perhaps higher standards established locally.

Various educational and policy problems are posed by these alternatives. For example, a single statewide standard for all students in a particular vocational program may be seen as both too difficult and too easy given wide variations in student ability and performance and, perhaps, in the varying demands of the marketplace. Differential standards require that each student's capacity be estimated, with the dual problems of the possible subjectivity of such estimates and the self-fulfilling prophecy phenomenon.

Moreover, if testing were designed to certify that students had achieved adequate proficiency to perform in the marketplace, such differentiation would deprive the certification of uniform meaning even at the lower end of the scale.

These sorts of educational and policy problems have legal analogs. A single statewide proficiency standard could be challenged on a number of grounds. If it failed to relate adequately to the demands of the job market, it could be challenged for lack of conformity with the state's educational quality responsibilities, or for its arbitrariness under due process notions. If the consequence of a single statewide proficiency standard had a sharply different impact on groups of students, especially those defined by race or ethnicity, an equal protection challenge might be forthcoming.

Resorting to multiple standards would not necessarily eliminate these legal concerns. Illustratively, if performance expectations for minority students were consistently and substantially reduced, although those students might be "certified," such an approach could stigmatize them, lower the program's expectations for them, and deny them access to remedial programs designed to elevate their proficiency levels. The consequence of these factors might actually be to diminish the job prospects of minority graduates of vocational education programs.

Differential standards could also raise substantial due process issues regarding the arbitrariness or irrationality of the standards themselves and of the mechanism by which they were set. The strength of this challenge would depend upon the care exercised by the responsible education authorities. If, for example, standards were established for each pupil by

an individual teacher acting impressionistically rather than on the basis of carefully articulated criteria, the system would be very vulnerable.

The level at which proficiency standards are set. A question related also to the number of standards set is whether standards ostensibly established to reflect the demands of the marketplace are for entry or journeymen-level positions. As a practical matter, unless a particular program is specifically designed to equip its students for journeymen positions, the standards should be geared to entry level positions. The more important issue is likely to be whether the standards actually relate to the marketplace.

There is evidence that in many vocational programs, instruction may not be effectively geared to the job market.[54] If that tendency were extended to performance testing standards,[55] there would be clear policy and legal problems. The performance testing effort could be attacked on the due process ground that it was not rationally related to the state's avowed purpose of equipping students to compete in the job market. Moreover, if the level at which standards were set did not comport with the marketplace, a state education clause challenge might lie. Finally, standard setting raises the issue of who makes the operative decision. It is inconceivable that standards could reasonably relate to the demands of the job market without the standard-setting process substantially involving representatives of the market in question. Nonetheless, from a legal perspective, the ultimate decision must be made by the responsible public officials. Otherwise, the standards are subject to challenge on the basis of an unlawful delegation of authority.

Whether the standards will be for school programs or for students. Thus far, this paper has proceeded primarily on the assumption that performance testing standards will be established for students rather than for school programs. This orientation is not inevitable. A performance testing program might be established to determine how well vocational schools or programs are performing on the whole.

The practical differences between these two approaches are substantial. As Brickell pointed out in connection with minimum competency testing, the choice between them will determine:

> ". . . whether you will write test items all students can pass or only most students can pass; whether you will test everybody or only a sample; whether you will report results to each individual parent or only to the general public; whether you will settle for a school program that reaches 70% of the students even if that 70% misses, for example, every single 'disadvantaged' child; and whether you will modify every unsatisfactory program or fail and recycle every unsatisfactory graduate."[56]

A focus on schools and their programs will reduce some legal difficulties but may increase others. To the extent that such a focus would reduce or eliminate sanctions against individual students or groups of students (i.e,. by not denying them promotion, graduation, or regular diplomas, or by not publicly identifying them as below proficiency levels), due process and equal protection concerns would be lessened. Arguments based on deprivation of a liberty or property interest, or on invidious discrimination, would be far less credible. The thrust of performance testing would be on school or program accountability and the response to inadequate performance presumably would be a programmatic or personnel-oriented response.

That may be a rational and appropriate approach unless the state's constitution, statutes or regulations impose a clear educational quality requirement directed to the rights of each student. In that event, as previously discussed, a performance testing effort, which was not designed to

ensure that each student had an educational opportunity geared to the achievement of reasonable proficiency in job-related vocational skills, would be suspect. Failure of the program to lead to special educational assistance for individual students who fell below the specified standards would be the clearest indication of its invalidity.

The consequences of failing to achieve the standards. This final keynote follows directly from the prior discussion. In connection with minimum competency testing, Brickell suggested six possible consequences for students who fell below minimum competencies and six parallel consequences for schools whose students failed to perform adequately. They were:

1. Verify the findings independently
2. Provide several more chances
3. Lower the standards to meet their performance
4. Remediate so that they can pass (or redesign school programs to match successful programs).
5. Refuse to promote or graduate them (or refuse to let schools operate until they can meet the standards)
6. Promote or graduate them with a restricted diploma or certificate or attendance (or let schools operate but refuse to accredit them.)[57]

In applying these possibilities to performance testing in vocational education, the prior discussion made clear that the preferable, and in some states the required, response to evidence that particular students have failed to meet proficiency standards is to direct appropriate educational assistance to them. This may take the form of remediation for the individual students; it may also involve broader programmatic or personnel responses. Surely if a substantial percentage of the school's or program's students is failing to meet statewide or local standards, the overall educational program, including the quality of instructional staff, should be evaluated and perhaps upgraded.

Lowering the performance testing standards because "too many" students have failed to meet them[58] is an unacceptable response for both public policy and legal reasons.

If students who fail to meet the standards are provided with appropriate remedial assistance and if the program is otherwise fair and rational,[59] then ultimately they could be refused promotion or graduation, or be promoted or graduated with a restricted diploma or certificate of attendance. From a due process perspective, these students may have been deprived of a liberty or property interest by that action but the state is permitted to do so if it acts fairly and rationally. From an educational quality perspective, the state cannot be required to guarantee educational results for all students. It can be held, however, to provide an appropriate educational opportunity for all students.

Vocational educational results, as measured by an effective performance testing program, are relevant to a determination of whether that educational opportunity is appropriate. In legal terms, evidence of inadequate pupil performance should shift to the education authorities the burden of demonstrating that, nonetheless, they have been providing their students with appropriate educational opportunities. This result is consistent with sound public policy and with the discharge by educators of their professional responsibilities.

Future Developments

The minimum competency movement has generated extensive debate and controversy. Its future is uncertain. Part of the uncertainty arises because pending and future legal challenges may invalidate entire programs or certain aspects of them. The *Debra P.* decision, the first relating to a direct minimum competency challenge, has not resolved the matter; indeed, it may have heightened the uncertainty by providing ostensible support for both supporters and challengers of minimum competency testing.

Uncertainty about minimum competency testing extends beyond the legal arena, however. Educators and policy makers are divided about the likely effects of these efforts. Whether the movement will improve education and educational outcomes by promoting more responsible and effective teaching, administering, and studying, or will victimize those who are held accountable by it, cannot be determined yet. In substantial part, the answer to that crucial question will turn upon the quality of further policy making that can shape or reshape minimum competency programs. It will also depend upon the care and skill exercised in implementing the policy thrusts.

The evolution of performance testing in vocational education hopefully should benefit from this experience in minimum competency testing. There are sufficient parallels to make this a reasonable possibility. What is required of policy makers and practitioners in vocational education is that they neither uncritically adopt performance testing as a solution to all their problems, nor reject it out of hand because it will have to be developed and implemented with thoughtfulness and care.

Legal principles, and the threat or actuality of litigation, may come to play an important role in the evolution of performance testing programs, too. This role, it is hoped, will be a positive one, requiring rationality, fairness and objectivity of the process, but not making impossible demands. But, vocational educators should not simply sit back and wait to be sued. They should deal in some preventive maintenance—they should attempt to head off legal challenges by fashioning and implementing performance testing programs in the most careful manner possible. If they do so, the law and the courts will have been an important partner in educational and professional reform.

Notes

[1]See, e.g., Ralph D. Turlington, "Good News from Florida: Our Minimum Competency Program is Working," *Phi Delta Kappan* 60 (May 1979): pp. 649-51.

[2]See, e.g., Merle S. McClung, "Competency Testing Programs: Legal and Educational Issues," *Fordham L. Review*, 47 (1979): 651-712. Donald W. Lewis, "Certifying Functional Literacy: Competency Testing and Implications for Due Process and Equal Educational Opportunity," *Journal of Law and Education*, 8 (April 1979): 145-83.

[3]See Chris Pipho, "Minimum Competency Will Disappear But Other Controls will Remain," *Phi Delta Kappan* 60 (February 1979): p. 412.

[4]Shirley B. Neil, "A Summary of Issues in the Minimum Competency Movement," *Phi Delta Kappan* 60 (February 1979): pp. 452-53.

[5]*Debra P. v. Turlington*, 474 F. Supp. 244 (M.D. Fla. 1979). Another minimum competency challenge, *Green v. Hunt*, Civil No. 78-539-Civ.-5 (E.D. N. Car. April 4, 1979), was dismissed for procedural reasons.

[6]*Hernandez v. Board of Education, Lynwood Unified School District*, Case No. SCC 01531 (Super. Ct. Los Angeles Co., filed May 22, 1979); *Wells v. Banks*, Civil No. CV 478-138 (S.D. Ga., filed June 17, 1978).

[7]See, e.g., Paul L. Tractenberg and Elaine Jacoby, "Pupil Testing: A Legal View," *Phi Delta Kappan* 59 (Dec. 1977): pp. 249-54.

[8]E.g., *James v. Board of Education, City of New York*, 42 N.Y. 2d 357, 397 N.Y.S. 2d 934, 366 N.E. 2d 1291 (Ct. App. 1977) (court refuses to intervene in determining whether integrity of citywide reading tests had been compromised).

[9]E.g., *Chappell v. Commissioner of Education*, 135 N.J. Super. 565, 343 A. 2d 811 (App. Div. 1975) (court refuses to intervene in "fundamental educational policy" decisions to initiate pupil testing and to disseminate results).

[10]E.g., *Larry P. v. Riles*, 343 F. Supp. 1306 (N.D. Cal. 1972), aff'd, 502 F. 2d 963 (9th Cir. 1974) (cultural bias of I.Q. tests against black children). The court recently reaffirmed that decision.

[11]Henry M. Brickell, "Seven Key Notes on Minimum Competency Testing," *Phi Delta Kappan* 59 (May 1978): pp. 589-92.

[12]J. Stanley Ahmann, "Implications of the Minimum Competency Testing Movement for Performance Testing in Vocational Education." An unpublished paper, 1979.

[13]The term "state" includes not only state government but also other state and local governmental bodies, including school districts.

[14]The U.S. Supreme Court, until 1937, struck down numerous state laws as not having a "real and substantial relationship" to permissible state purposes and, therefore, as being violative of "substantive due process." The most notorious example of such activity was *Lochner v. New York*, 198 U.S. 45 (1905), where the Court struck down New York's maximum hour legislation for bakery employees. The breakthrough case where the Court applied the now common and more relaxed "rational basis" test was *West Coast Hotel Co. v. Parrish*, 300 U.S. 379 (1937). Since that time, other than in cases dealing with civil rights and civil liberties, virtually no state law has been invalidated by the Supreme Court as being violative of "substantive due process".

[15]Merle S. McClung, "Competency Testing: Potential for Discrimination," *Clearinghouse Review* 11 (1977): pp. 439-48.

[16]E.g., *Goss v. Lopez*, 419 U.S. 565 (1975).

[17]Stigmatization, as infringing on a protected liberty interest, was recognized in *Wisconsin v. Constantineau*, 400 U.S. 433 (1971). See also *Board of Regents v. Roth*, 408 U.S. 564 (1972). Paul L.Tractenberg, "Selecting 'Educationally Deprived' Students for Title I: A Review of the Legal Issues," (an unpublished paper prepared for the National Institute of Education, 1977): 59-62. However, in *Paul v. Davis*, 424 U.S. 693 (1976), the United States Supreme Court narrowed the definition of stigmatization to require the "alteration of legal status which, combined with the injury from defamation, justified the invocation of procedural safeguards." 424 U.S. at 708-09.

[18]See, e.g., *Rizzo v. Goode*, 423 U.S. 362 (1976); William J. Brennan, "Address to the New Jersey Bar," May 22, 1976 [reprinted in Guild Practitioner 33 (1976): pp. 152-68.]

[19]See, e.g., *People v. Brisendine*, 13 Cal. 3d 528, 531 P. 2d 1099, 119 Cal. Rptr. 315 (1975).

[20]For a comptency test to meet technical requirements, it must be shown that it is both valid and reliable. Validity refers to whether the test actually measures the characteristic that it claims to measure. Reliability refers to whether the test measures that characteristic accurately and consistently. In the case of competency testing an invalid reading test might actually be measuring writing skills. An unreliable reading test might give a student who took the test twice, using two different forms of it, a high score when he or she used form A and a low score when he or she used form B. See American Psychological Association, *Standards for Educational and Psychological Tests* (1974).

[21]This also could be the basis for a due process challenge—namely, that the state was acting irrationally. See, e.g., Arthur Wise, "Minimum Educational Adequacy: Beyond School Finance Reform," *Journal of Education Finance 1* (Spring 1976): 468-83; Joan Baratz, "In Setting Minimal Standards Have We Abandoned Concerns for Equity and Access," Paper presented at Wingspread Conference, Educational Policy Research Institute, Washington, D.C., July 1978. See also *Phi Delta Kappan* 59 (May 1979), which contains a series of articles on minimum competency testing.

[22]*San Antonio Independent School Dist. v. Rodriquez*, 411 U.S. 1 (1973).

[23]See, e.g., *Serrano v. Priest*, 18 Cal. 3d 728, 557 P. 2d 929, 135 Cal. Rptr. (1977); *Horton v. Meskill*, 172 Conn. 615, 376 A. 2d 359 (1977).

[24]In *Washington v. Davis*, 426 U.S. 229 (1976), the Court held that disproportionate racial impact of a test is insufficient to establish an unconstitutional racial classification; a discriminatory purpose must be shown. Several subsequent Supreme Court decisions shed light on how that purpose may be shown. See, e.g., *Village of Arlington Heights v. Metropolitan Housing Development Corp.*, 429 U.S. 252 (1977). In light of this narrowing construction of the equal protection clause, challenges based upon Title VI of the Civil Rights Act of 1964 and its implementing regulations may be preferable. The U.S. Supreme Court indicated in *Washington v. Davis* that disproportionate racial impact of a test might be sufficient to constitute violation of Title VI. See McClung, supra, n. 15, at 442.

[25]See, e.g., *Swann v. Charlotte-Mecklenburg Board of Education*, 402 U.S. 1 (1971).

[26]See, e.g., *Wisconsin v. Yoder*, 406 U.S. 205 (1972); *Tinker v. Des Moines Independent School District*, 393 U.S. 502 (1969); *West Virginia State Board of Education v. Barnette*, 319 U.S. 624 (1943). See generally McClung, supra n. 2, at 674-77.

[27]The education clauses use a variety of formulations. Among the more common descriptions of the required educational quality are the following: (i) "thorough and efficient" (e.g., N.J. Const. art. VIII, §4, 1; Ohio Const. art. VI, §2; Pa. Const. art. III, §14, W. Va. Const. art. XII, § 1); (ii) "high quality" (e.g., Ill. Const. art. X, §1; Mont. Const. Art. X §1(3); Va. Const. art. VIII, §1); (iii) "general and uniform" (e.g., Ariz. Const. art. XI, §1; Idaho Const. art. IX, §1; Ind. Const. art. VIII, §1).

[28]See Paul L. Tractenberg, "Legal Implications of Statewide Pupil Performance Standards." Paper prepared for the Education Commission of the States, September 1977.

[29]In *Robinson v. Cahill*, 62 N.J. 473, 303 A.2d 273 (1973), the New Jersey Supreme Court interpreted the state's "thorough and efficient" clause in that manner.

[30]This is likely to be the most difficult link to establish. A performance testing program undeniably is a rational way for the state to implement its educational obligation. But the state will maintain that there are other rational ways available to it.

[31]This approach would raise formidable proof problems and the challengers would have to overcome a court's tendency to defer to the expertise of legislators or educators who have set the standards.

[32]See n. 27 supra.

[33]*Hernandez v. Board of Education, Lynwood Unified School District*, supra. n. 6.

[34]E.g., P.L. 94-482, §112 (1976).

[35]E.g., Title VI of the Civil Rights Act of 1964, 42 U.S.C. §2000d (1976); Equal Educational Opportunity Act of 1974, 20 U.S.C. §§1701-1758 (1976).

[36]Family Educational Rights and Privacy Act of 1974, 20 U.S.C. §1232(g) (1976), P.L. 90-247, as added P.L. 93-3800 and amended P.L. 93-568. Implementing regulations are at 45 C.F.R. 99.1 et seq.

[37]E.g., N.J.S.A. 18A:7A-2(a) (5), (6), (7).

[38]E.g., Education for all Handicapped Children Act of 1975 (P.L. 94-142), 20 U.S.C. §§1401-1461 (1976). Implementing regulations are at 45 C.F.R. §121a.1-754. (1978).

[39]Educational malpractice cases are probably the best known lawsuits regarding pupil performance. Those cases are based primarily on common law negligence theories. The assertion is that students have failed to learn because the schools and their professional staffs have breached a duty of care and skill owed to the students. Thus far, educational malpractice cases on behalf of "normal" students have been unsuccessful because of the courts' public policy concerns about imposing such liability on school systems and professionals. See, e.g., *Peter W. v. San Francisco Unified School District*, 460 Cal. App. 3d 814, 131 Cal. Rptr. 854 (Ct. App. 1976); *Donohue v. Copiague School District*, 64 A.D. 2d 29, 407 N.Y.S. 2d 375, 391 N.E. 2d 1352 (Ct. App. 1979). Cases brought on behalf of handicapped students alleging particular negligent acts of specified professionals, rather than a general pattern of negligence, have been more successful. See, e.g., *Hoffman v. Board of Education, City of New York*, 64 A.D. 2d 369, 410 N.Y.S. 2d 99 (App. Div. 1978). Recently, however, the New York Court of Appeals reversed the *Hoffman* decision on public policy grounds. Although the results of performance testing in vocational education might highlight inadequate performance of some students, those results are unlikely to cause the judiciary to depart substantially from the policy approach it has staked out. See generally Note, "Implications of Minimum Competency Legislation: A Legal Duty of Care," *Pac. Law Journal* 10 (1979): 947-70.

[40]See Ahmann, supra, n. 12, at 8-11.

[41]See, e.g., *Robinson v. Cahill*, 62 N.J. 473, 303 A. 2d 273 (1973).

[42]Validity has both a generalized meaning of suitability and appropriateness, and a technical psychometric meaning. As to the latter, see n. 20 supra.

[43]See n. 20 supra.

[44]See n. 17 supra.

[45]See n. 16 supra.

[46]See n. 24 supra.

[47]In *Village of Arlington Heights v. Metropolitan Housing Development Corp.*, 429 U.S. 252 (1977), the Court listed a number of factors that may be considered in establishing discriminatory intent. These included: (1) historical background; (2) the specific sequence of events leading up to the challenged decision; (3) the departures from normal procedural sequences or typical substantive results; and (4) the legislative or administrative history.

[48]See, e.g., *Ingraham v. Wright*, 430 U.S. 651 (1977); *Rizzo v. Goode*, 426 U.S 362 (1976). See also Tractenberg, supra n. 7, at 13.

[49]474 F. Supp. n. 23, at 261.

[50]See Ahmann, supra n. 12, at 19.

[51]The focus of this performance testing probably will be whether the student has adequately mastered certain foundation or prerequisite skills.

[52]Diana Pullin's paper deals with this issue in more detail.

[53]The court in *Debra P.* deferred the Florida diploma sanction for four years because students had not had adequate notice of, or opportunity to prepare for, the competency test. The court also addressed the unavailability of meaningful remedial programs until shortly before the sanction attached.

[54]See *New York Times*, October 16, 1979, §C, at 1, Col. I.

[55]If the performance testing standards were related to the marketplace but the instruction actually provided in the program was not, there would be a mismatch between course and test content. This would raise issues of substantive due process.

[56]See Brickell, "Seven Key Notes," p. 592.

[57]Ibid.

[58]Evaluation instruments, and perhaps the performance testing standards themselves, can be modified if, based on field testing or otherwise, valid educational or psychometric judgments indicate that modification is required to implement the state's goals. Safeguards should be erected, however, to prevent this from being an open door to dilution of standards. If standards were lowered so that they no longer were reasonably related to the demands of citizenship and the job market, they could be challenged on legal theories discussed previously.

[59]Some of the primary elements of a fair and rational system are: (i) carefully developed, non-discriminatory standards; (ii) valid evaluation instruments and procedures; (iii) an opportunity for verification of the initial evaluation results; and (iv) evaluation early enough to permit remedial assistance (or program redesign) and re-evaluation. Some commentators have also suggested that testing programs should be phased in so that students who have substantially completed the educational process do not have new and onerous standards imposed upon them. See McClung, "Competency Testing," p. 2.

Accountability Through Performance Testing

Performance testing instruments and techniques are designed to foster accountability within the vocational education system. The use of performance testing for accountability raises legal concerns on behalf of both students and the vocational educators conducting the testing program. In both cases, the importance of the legal considerations will be directly related to the extent of the harm resulting from the use of the tests. In some instances, the legal issues for educators and for students will overlap.

While this paper will focus for the most part on legal issues arising from harm to students from a performance testing program, it is helpful to enumerate the legal impact on educators themselves. Performance testing is initiated to foster accountability in vocational education, but that accountability can be designed to diagnose weakness and provide effective feedback for change or to diagnose weakness and eliminate that weakness. Results of student performance tests can be used to evaluate and guide teacher or program effectiveness, or tests can be used to assist teacher termination decisions. The former situation raises few legal issues; the latter presents issues that have been addressed previously by the judiciary.

The most striking example of the use of student tests for teacher accountability involved the termination of an elementary teacher due, in large part, to the performance of her students on standardized achievement tests, the Iowa Tests of Basic Skills and the Iowa Tests of Educational Development. While the trial court found that the dismissed teacher should be reinstated, an appellate court disagreed and upheld the teacher's dismissal. The appellate court noted a dispute among educators about the reasonableness of using the tests to assess teacher competence but found that the action of the school board and the superintendent in the dismissal was reasonable.[1] It is not unreasonable to expect that a court might have the same reaction to the use of vocational performance tests for teacher termination.

A court's analysis of the legality of the use of performance testing to evaluate and terminate teachers rests in large part upon an examination of whether the scheme complies with constitutional guarantees of due process of law or fundamental fairness.

Fundamental Fairness

An area where educational and legal policy questions most closely coincide concerns the fundamental fairness of performance testing programs. Within the legal system, this issue is addressed by assessing whether the program meets constitutional standards of due process of law. This issue is addressed by assessing whether the testing program is designed to serve a necessary and legitimate governmental purpose and is formulated to serve that purpose through reasonable means. Within the educational system, this issue is addressed by assessing whether a testing program serves the educational goals and objectives of the schools.

Traditionally, constitutional guarantees of due process of law insure that individuals are treated with fairness, consistency, and lack of arbitrariness by governmental agencies and employees. Due process protections are of two types: procedural and substantive. Procedural due process protections seek to insure that the procedures used by government in dealing with individuals are fair. Procedural due process protections typically include the right to some form of notification of impending governmental action and the right to effectively influence or participate in governmental decision-making through hearings, representation by counsel, review of evidence, and so forth. Substantive due process seeks to ensure that, regardless of the

procedures followed, the action undertaken by the government must be reasonable and must serve a legitimate governmental objective or purpose.

Due process, both substantive and procedural, is an elastic concept requiring different levels of protection depending on the context. The procedural protections that must be afforded a defendant in a criminal trial are much more detailed than those that must be provided a student who faces a long-term suspension from school. Similarly, the governmental objectives to be served by a statute regulating conduct through criminal sanctions will be subject to a much stricter substantive due process analysis than the objectives of a statute regulating the dress and appearance of police officers. While the meaning of due process, the delineation between substantive and procedural due process, and the standards for determining what process is due in a particular situation can be somewhat blurred. However, there are guidelines offered educational decision-makers and vocational educators by a due process analysis of performance testing schemes.

A substantive due process analysis ordinarily begins with an examination of the legitimacy of the goal of the governmental program. This analysis of the "state interest" in a program can rarely be conducted by referring to a full and clearly articulated statement by the governmental agency made at the time the program was initiated; such statements seldom exist. Instead, a court relies upon the government's after-the-fact rationale for its program or the court itself defines what it feels a legitimate interest or goal might be. A substantive due process analysis therefore begins with scrutiny of the goals, either explicit or implied, of a testing program. Next, if the governmental goals are legitimate (and courts almost always find that they are), the means of achieving the goal will be examined.

Two examples of judges' use of substantive due process to analyze educational practices may be helpful. Both situations involved school discipline and the exclusion of students from school for alleged violations of school rules of conduct. In the first case,[2] a New Hampshire high school student was indefinitely expelled from school for intoxication. Laws of the State of New Hampshire permitted expulsion of students for "gross misconduct;" school rules specified that students could be expelled for "undesirable behavior patterns." The expelled student's infraction of the rules was her first offense, there was no evidence of any disruption of other students, and evidence presented to the judge hearing the case indicated that the misbehavior was due in large part to difficulties that student had been having in her relationship with her parents. In the New Hampshire case, the court stated that:

> It is fundamentally unfair to keep a student out of school because of difficulties between the student and her parents, unless those difficulties manifest themselves in a real threat to school discipline.[3]

In reaching a decision which ordered the student reinstated in school, the court considered the harm to the student in being excluded from school, the effectiveness of the exclusion in deterring other student misconduct, and the failure of the school to prove that readmitting the girl to school would cause significant harm to the school's functioning. In addition, the analysis focused upon whether it is fair to punish students for behavior over which the students themselves have little, or no control.

In the second case,[4] a brother and sister were both suspended from a Louisiana school under a school rule which allowed for the discipline of a student when the student's parent challenged the authority of school officials in an "offensive manner." The students were suspended indefinitely and then transferred to a new school for disciplinary reasons after their mother struck an assistant principal in the course of a discussion over his discipline of the

children. A federal court of appeals found the discipline of the students an unconstitutional infringement of the right to substantive due process of law. The school rule punished students in the absence of any personal guilt for the infraction and in a situation where the school could not meet a substantial burden placed on it to justify its actions. The Louisiana case involved a similar analysis. There, the court asked whether there was a justifiable and reasonable need for the school rule punishing students for the misconduct of their parents and whether there was a reasonable and less onerous alternative means for fulfilling the need the rule was designed to serve.

A second series of questions relating to the fundamental fairness of an educational program or practice concerns the manner in which the program or practice was implemented. These questions are sometimes treated as procedural due process issues, sometimes as substantive due process issues. The implementation of a new program or practice presents fairness issues which relate both to the sufficiency of advance notice of the change (procedural due process) and to the extent to which the implementation scheme reasonably and rationally furthers a legitimate educational purpose (substantive due process). Because a procedural due process analysis is most often applied to situations scrutinizing the mechanics of formal or informal procedures involving hearings, the substantive due process rubric may be more helpful here.

One court was asked to apply a due process analysis to a situation in which a student challenged the manner in which her graduate program changed the requirements for a master's degree. In that case,[5] the student argued that she was denied procedural and substantive due process guarantees when a comprehensive examination was added as a graduation requirement after she had commenced her graduate program. The appellate court considering the case decided in favor of the school after analyzing the factors involved. The court, however, implicitly recognized a due process right to timely notice of a change in graduation requirements.

There is clearly a legitimate governmental interest in maintaining decorum in the schools through school discipline rules. The substantive due process considerations presented in the two cases described above concern whether the school rules were fair means of achieving that goal and whether the rules were fairly applied. A similar type of due process analysis to that described in the two discipline cases can be followed in examining school testing programs. The analysis has already been applied to the statewide use of a minimum competency testing program to deny high school diplomas.

The Fundamental Fairness Flaw In One Minimum Competency Testing Program

A forecast of the type of substantive due process analysis that might be applied to performance testing in vocational education can be formulated by examining a recent court decision concerning Florida's use of minimum competency testing to deny high school diplomas. The decision was made in the case of *Debra P. v. Turlington*[6] in the summer of 1979 and was the first judicial reaction to the legality of the minimum competency testing movement then sweeping the nation's secondary schools. The lawsuit was brought by a number of students who failed the competency test and would, as a result, be denied regular high school diplomas and awarded instead certificates of completion of high school.

Florida's minimum competency test requirement was the result of a 1976 state law concerning educational accountability. The law required that high school graduates be provided at least the minimum skills necessary to function and survive in modern society and that students demonstrate satisfactory performance in "functional literacy" to receive a high school diploma.

Pursuant to the statute, a minimum competency examination of functional literacy was administered to Florida's public high school juniors and seniors. The functional literacy test was first given to juniors in the fall of 1977; students who failed the test had two more chances to take it before the graduation requirement was to be imposed in the spring of 1979. Substantial numbers of stuents and a 'disproportionate number of black students,' failed the test.

The students who brought the lawsuit challenging the Florida testing program based their challenge on several different claims: that the program resulted in unlawful racial discrimination; that the program, through the remedial classes provided to students who failed the test, resulted in resegregation of black students, and that the program denied due process of law. After a lengthy trial, the court issued a decision that placed a four-year moratorium on the use of the functional literacy test to deny high school diplomas.

The substantive due process analysis is of primary interest as an analogy for studying performance testing. The Florida court had little difficulty in finding a legitimate purpose served by the testing, i.e., ". . . the test could be utilized not only to gauge achievement, but also to identify deficiencies for the purpose of remediation."[8] The issue of the legitimacy of the means used to reach this goal was of greater difficulty. The issue, as the court saw it, was ". . . whether the test utilized was a valid and reasonable measure for dividing students into classifications for the purpose of high school graduation."[9] One might well ask whether the court was confused about what the goals and means involved were. The due process issue which had been presented to the court was whether the test instrument itself and the means used to implement the testing program were fair means to achieve the goals of placing students in remedial classes, label test failers as "functional illiterates," and to determine the award of high school diplomas in lieu of certificates of completion.

A major criteria for review of the testing program concerned whether adequate notice of the change in the graduation requirement was provided to parents, students, and educators. Florida's statute was passed in the summer of 1976; the standards and objectives to be measured on the test were established in the spring and summer of 1977; the first functional literacy examination was administered in the fall of 1977. In effect, teachers in Florida's high school had only two months of class time to work with students on the new functional literacy skills measured on the test, skills which the court found had not previously been successfully taught to all of Florida's students.

The court recognized the need to inform students and educators of the importance of the test and the sanctions to be imposed as a result of the test, in addition to the subject matter to be examined. The court recognized the educational implications of adequate notice:

> While all instruction is important, there are obvious methods of motivating students and emphasizing certain skills. The principal problem with the instant program is that the instruction in previous years took place in an atmosphere without the diploma sanction . . . It is critical that at the time of instruction of a functional literacy skill, the student knows that the individual skill that is being taught must be learned prior to his graduation from a Florida public high school. Instruction in the specific skills is critical, but likewise so is identification of whether the skills have been learned. Teaching and learning are not always coterminous.[10]

Based upon the expert testimony of several educators, the court concluded that four to six years should intervene between the time the objectives to be measured on the test are made public and the sanction resulting from the test is implemented.

To assess the validity, reasonableness, and arbitrariness of the minimum competency test, the court discussed the content and construct validity of the test and alluded to other technical flaws in the test development and administration process. The court noted errors of "considerable magnitude" in test development and administration and found adequate levels of content and construct validity. The court found that, even if Florida's test developers did not meet appropriate professional standards of "state of the art" requirements, constitutional due process standards are not identical to professional test and measurment standards. A test, according to the Florida court, need only bear a rational relation to a valid state interest. The constitutional standards for test instruments themselves are therefore lower, in certain cases, than professional standards.

Fundamental Fairness In Performance Testing For Vocational Education—Some Recommendations

In the context of performance testing in vocational education, what due process is due? Clearly, for those programs where successful test performance is required to exit from the program, obtain a certificate or license, or for entry into an apprenticeship following the formal training, students should be fully informed of the test requirement before entering the program. The nature of the sanction, e.g., failure to complete a course of study, to obtain a license or certificate, or failure to be apprenticed, is of sufficient magnitude to require the early and complete notice. What of tests of less magnitude? Given the reluctance of the judiciary to become involved in educational decision-making, particularly in individual relationships between instructors and students,[11] a court may never intervene to determine the degree of due process appropriate for such a situation. Court intervention, and the extent of such intervention, will always hinge on the extent of the harm resulting from a program or practice. However, the basic tenets of due process would indicate that, if imposed, the due process requirements are less strict, that the notice can be less complete when tests have less importance. An instructor in an occupational home economics course giving a test at the end of a teaching unit on metric conversion would, for example, be held to far less strict requirements than was the State of Florida in testing to deny high school diplomas.

The nature of judicial involvment to one side, would it not be appropriate for educators to impose some due process, or fundamental fairness, requirements upon themselves in the classroom testing situation? Such requirements would undoubtedly foster better teaching and more effective learning, Educators have recognized the importance of careful objective setting for both teacher and learner.[12] There should be little dispute that vocational students would benefit from knowing in advance what is to be expected of them as a result of their training, and that learning will improve as goals are clearly identified and worked toward. Any constitutional due process standard of notice that would be applicable in this situation would not impose an additional requirement on educators but would instead simply restate the perimeters of good educational practice.

Assuming that a performance testing requirement has been fairly imposed, some guidelines concerning the nature of the test itself can also be drawn from the Florida court's reaction to the high school minimum competency test. What technical standards of the test and measurement profession have been recognized by the judiciary as applicable to educational testing?

The Florida court, in its discussion of due process notice requirements was, in effect, recognizing the seldom recognized but increasingly important concepts of curricular and instructional validity. There was, in short, no match between the functional literacy skills and objectives measured on the minimum competency test and the curriculum and instruction offered the students who were required to pass the test to receive a high school diploma.

To achieve fundamental fairness in performance testing for vocational educators, schools and instructors administering the tests should follow the following guidelines:

- Students should be informed of the existence and the nature of the testing requirement well in advance of taking the test.
    - If the performance test will be required to exit or graduate from the training program, the student should be informed of the test before entering the program.
    - If the performance test will be required to complete a course or unit of study successfully, the student should be informed of the test before beginning the course or unit.
- Students should be informed of the subject-matter, skills, and objectives to be measured by the test.
- The curriculum and instruction offered the student should cover all subjects, skills, and objectives to be measured by the test.
- The test should only measure those areas actually covered by curriculum and instruction.
- The test instruments of techniques used should meet professional standards for validity and realiability.

Performance Testing and The Potential For Unlawful Discrimination

In addition to the fundamental fairness issues addressed by the Florida court considering minimum competency testing, the court also addressed issues of unlawful racial discrimination resulting from use of the functional literacy test. Similar issues are presented by performance testing in vocational education.

Florida's functional literacy test, after the third administration just prior to graduation, had failure rates that clearly indicated that the testing program impacted disproportionately on black students. The failure rate for black students was approximately ten times that among white students. The students challenging the test alleged that the test results for black students reflected the educational deprivations those students had suffered; the high school seniors who faced the test-for-graduation requirement spent the crucial first four years of their schooling in inferior, racially segregated schools. In the years since physical integration of the schools, black students continued to suffer ongoing discrimination. Poor test performance for black students both reflected and perpetuated the effects of past racial discrimination.

The judge, considering these arguments against the Florida test, determined that the use of the functional literacy test to deny high school diplomas constituted unlawful racial discrimination. The test, the court concluded, should not be used as a graduation requirement until all of the seniors compelled to meet the test-for-graduation requirement had completed a full twelve years of physically desegregated schools. Thus, a four-year moratorium on the use of the test as a graduation requirement was ordered.

The race discrimination analysis in the Florida case was based upon both a constitutional and a statutory theory. Under the constitution, the testing program denied equal protection of the laws to black students. Under federal statutes, the program violated Title VI of the Civil

Rights Act of 1964. The use of constitutional and Title VI theories to scrutinize an educational testing practice has been employed recently, with perhaps even more far reaching implications than the Florida case, by a federal court in California. The California case[13] involved the use of I.Q. tests to place students in classes for the educable mentally retarded (EMR). Classes for EMR students were populated with a large percentage of black students, a percentage considerably higher than the proportion of blacks in the total school population. The court found it unlawful to rely upon I.Q. tests to determine EMR placement when there is no proof that those tests are valid and reliable for use with black students and there is no proof that use of the tests or resultant disproportionate class placements furthered the purpose of providing the best educational opportunities for students.

Challenges under both constitutional and Title VI theories could also be brought against performance testing in vocational education. There are also additional legal claims that can be brought in the vocational education context.

Programs that receive federal financial assistance are obligated to comply with an array of statutes and regulations prohibiting discrimination on the basis of race, sex, national origin, color, or handicap.[14] The nature of these prohibitions can be fairly summarized by reference to the March 21, 1979 "Guidelines for Eliminating Discrimination and Denial of Services on the Basis of Race, Color, National Origin, Sex, and Handicap" promulgated for vocational education by the Office of Civil Rights, Department of Health, Education, and Welfare. These vocational education guidelines set forth nondiscrimination requirements concerning distribution of funds, access and admissions to programs, counseling and prevocational programs, instructional programs, employment of faculty and staff, and proprietary schools.

The vocational education guidelines set forth several standards relevant to performance testing situations. The guidelines indicate that programs may not develop, impose, maintain, approve, or implement discriminatory admissions criteria. Programs may not ordinarily judge candidates for admission on the basis of selection criteria that have the effect of disproportionately excluding persons of a particular race, color, national origin, sex, or handicap. However, if a program can demonstrate that the criteria for admission have been validated as essential to participation in the program and that alternative and equally valid criteria without disproportionate impact do not exist, then the criteria may be used despite their disproportionate impact on the protected groups. A performance test measuring entry level skills used to select candidates for a vocational program could be subject to scrutiny under the guidelines. If the performance test used for admissions purposes resulted in a disproportionate number[15] of minority students failing the test, then the test could not be used unless vocational educators could demonstrate that the entry level skills being measured on the test were essential for successful participation in the vocational program. Even once this proof was made, the test could still not be used for admissions purposes unless there was no other valid way of assessing the entry level skills that did not have a disproportionate result.

The vocational education guidelines have similar types of nondiscrimination provisions that could also apply to uses of performance testing. For example, there can be no discrimination in making work-study or apprenticeship opportunities available to vocational students. Therefore, performance tests used to measure student readiness for a work experience should not result in disproportionate minority failure rates.

Nondiscriminatory Performance Testing-Some Recommendations

Legal standards regarding nondiscrimination are not designed to prohibit testing nor to circumvent the primary use of tests, e.g., discriminating between those who know or can perform from those who do not know or cannot perform. The legal standards being discussed here do, however, prohibit distinctions between test takers when the distinctions are based upon protected status, such as race, rather than upon knowledge or skill.

When the results of a test make it appear that distinctions were based upon race, national origin, color, or sex rather than upon true ability to perform the tasks being tested, then educators are asked to scrutinize their conduct to eliminate bias. This scrutiny has two phases: Does the test really measure something that has to be performed to succeed in the vocation for which the student is being trained, and is this test the only valid source of measurement or is there an alternative that will achieve the same goal without harming minorities?

In beginning a performance testing program, vocational educators can take the following steps to minimize the potential for unlawful discrimination:

- Test only at the basic level at which competence must be demonstrated; if a program is designed to produce apprentice plumbers, the performance test used for exit from the program should not measure skills that only a master plumber would be expected to know.
- If test results indicate that a disproportionate number of minority students are failing the test, determine whether there is a different but equally valid test that would measure the same areas, without the disproportionate result. Also, determine whether the test results reflect past deprivations and how these can be remedied through compensatory educational programs.

Performance Testing and the Right to Privacy

A final set of issues of legal concern relate to the use of performance test results once they are obtained. What use is made of the test results within the vocational program and how or where are performance test results disseminated outside the training program? For educational programs receiving federal financial assistance, there are clear standards concerning privacy and confidentiality under the Family Educational Rights and Privacy Act (FERPA).[16]

FERPA details protections for students concerning information, such as performance test results, contained in school records. Test results may not be disclosed to someone who does not have a "legitimate educational interest" in seeing the results without written consent from either the parent of the student or, for students over eighteen years old, the students themselves. Persons with "legitimate educational interests" and for whom consent is therefore unnecessary are probably only persons directly involved in the student's training program. Potential employers clearly should not receive such information without written consent; potential supervisors for a work-study or apprenticeship experience probably should not receive the information without written consent.

The federal student records law also requires that students and parents be provided interpretations of test result information should they request it. This provision clearly points to the need for careful and unbiased record keeping, the use of valid and defensible tests, and the need for trained personnel who can explain and counsel about performance testing.

In addition to the requirements of the federal records statute concerning privacy and consent, there are potential problems of a constitutional dimension concerning performance tests and the use of test results. Some educators may be inclined to include questions designed to assess student attitude about a task or vocation. Such attempts unlawfully infringe upon student privacy particularly when they scrutinize areas that are actually unrelated or unnecessary to successful performance in either the training program or the vocation. For example, a female student's attitude about pregnancy and child-rearing has no bearing on her potential as a secretary.

Privacy and Confidentiality in Performance Testing – Some Recommendations

To minimize potential infringement of students' privacy and the right to confidentiality, the following guidelines are appropriate:

- Test scores should not be disclosed to persons outside the school or to those not directly involved with the student's training without consent.
- Test scores should not be divulged to potential employers without the written consent of the parent or, where the student is over eighteen, the student.
- Interpretation of test results should be made available to students and parents
- Tests should not include questions that unnecessarily infringe on students' privacy.

Conclusion

The use of performance testing in vocational education can lead to desirable improvements in the delivery and outcome of training programs. Performance testing does present potential legal problems of some magnitude. None of these problems is insoluble and, in fact, a wise vocational educator will work to alleviate legal entanglements and will, at the same time, have improved the educational program.

To maximize the educational benefits of a performance testing program and to minimize the impact of legal scrutiny of the program, vocational educators should structure the program so that there is adequate phase-in time prior to implementation of the test. During the phase-in period, test developers should undertake efforts to insure the validity and reliabiltiy of the test instrument. During the phase-in period, instructors should inform students of the subject-matter, skills and objectives to be measured on the test and should insure that the areas covered on the test are in fact being taught all students. Next, educators and test developers should insure that tests do nt unlawfully discriminate against students on the basis of race, sex, national origin, or handicap. Finally, steps should be taken to protect the privacy of students participating in the testing program.

Notes

[1]*Scheelhaase v. Woodbury Central Community School District*, 349 F. Supp. 988 (N.D. Iowa, 1972), rev'd. 488 F. 2d 237 (5th Cir. 1973), cert. den. 94 S.Ct. 3173.

[2]*Cook v. Edwards*, 341 Supp. 307 (D.N.H. 1972).

[3]341 F. Supp. 309.

[4]*St. Ann v. Palisi*, 495 F. 2d 423 (5th Cir. 1974).

[5]*Mahavogsanan v. Hall*, 529 F. 2d 448 (5th Cir. 1976).

[6]474 F. Supp. 244 (M.D. Fla. 1979)

[7]A black student had a ten times greater chance of failing to graduate than did a white student.

[8]474 F. Supp. 260.

[9]Id.

[10]474 F. Supp. 264.

[11]The clearest example of this judicial reluctance is a recent U.S. Supreme Court case, *Board of Curators of the University of Missouri v. Horowitz*, 98 S. Ct. 948 (1978). In that case the Supreme Court noted, in discussing the academic expulsion of a medical student,that a student's academic status requires expert evaluation of cumulative information and a court should decline to overturn the judgment of educators.

[12]Robert F. Mager, *Preparing Instructional Objectives* (Belmont, Calif.: Fearon Publishers, 1975).

[13]*Larry P. v. Riles*, No. C-71-2270 RFP, N.D. Calif. October 10, 1979.

[14]Discrimination on the basis of race, color, or national origin is prohibited by Title VI of the Civil Rights Act of 1964, 42 U.S.C. §2000d. Discrimination on the basis of sex is prohibited by Title IX of the Education Amendments of 1972, 20 U.S.C. §§1681 *et seq*. Discrimination on the basis of handicap is prohibited by §504 of the Rehabilitation Act of 1973, 29 U.S.C. §794, and by P.L. 94-142, Education for all Handicapped Children Act, 20 U.S.C. §§1401 *et seq*. Each relevant statute has a set of implementing regulations, written by the U.S. Department of Health, Education and Welfare, to further clarify the law. Finally, Title II of the Education Amendments Act of 1976, 20 U.S.C. §§2301 *et seq*. and "Guidelines for Eliminating Discrimination," 44 Fed. Reg. 17162, referred to hereafter as "voc ed guidelines," also contain relevant nondiscrimination provisions.

[15]For the purposes of this discussion, a "disproportionate result" or "disproportionate effect" of a test is defined as a circumstance in which the total percentage, or proportion, of minority students failing the test is greater than that group's proportion in the total group of students taking the test.

[16]20 U.S.C. §1232g. The implementing regulations are found at 45 C.F.R. Part 99.

Comments on the Legal Issues in Performance Testing

William G. Buss
Iowa College of Law
Iowa City, Iowa

A hard lesson for law students to learn is that the expertise of a lawyer has much more to do with predicting legal outcomes than memorizing a set of rules. Making such predictions entails familiarity with the process of decision, appreciation of the distinct institutional roles of court and other decision-makers, and awareness of the constant interplay of fact determination and value judgment. Making such predictions employs a process of reasoning that is hardly scientific—in fact, a reasoning process that takes uncertainty as a pervasive feature of a dynamic system. Part of the lesson to be learned is that the predictions are characteristically tentative and often amount to little more than identification of alternative possibilities.

The truth contained in this lesson can be seen in the papers by Tractenberg and Pullin dealing with legal implications of performance testing for vocational education. These papers do not tell us—as they cannot—what legal results will follow from "performance testing;" they merely give tentative predictions—or, more accurately, they provide a legal framework within which predictions might be made. They tell us a little about the way courts work. For example, they make it clear that courts attempt to assimilate "real world" problems into legal categories, such as "due process of law" or the "equal protection of the laws" or a "right to privacy."

Tractenberg and Pullin tell us that the courts will both second guess educational judgments and defer to educational expertise, and they try to suggest when courts will do more of one and when more of the other. They tell us that the courts will examine facts—such as those provided in the testimony of educational experts or written in educational books or, perhaps, facts that are "known" by everyone, including judges, such as facts concerning the existence and disadvantage of racially segregated schools. They tell us also that the courts will make value judgments—such as those involved in, somehow, "weighing" the interests of individuals who may be harmed by denial of a high school diploma against the interests of the state in safeguarding the significance of a high school diploma.

Finally, Tractenberg and Pullin tell us that to hazard a prediction concerning the success of various legal challenges to performance testing one embark on a process of reasoning that is truly labyrinthian. For example, to predict the outcome of a discrimination challenge one must assess the intertwining significance of (a) certain Supreme Court cases dealing with the equal protection clause of the Fourteenth Amendment; (b) certain Supreme Court cases dealing with statutory provisions, such as Title VII of the Civil Rights Act of 1964, as amended (preventing employment discrimination); (c) a body of literature (not court decisions) dealing with competency testing (not, as such, performance testing in vocational education); (d) a single case by a court at the lowest level of the federal judicial system dealing with a particular competency testing law (again, not a law dealing with vocational education) in the particular context of a state educational system not yet freed from the constitutional implications of having had, prior to 1954, separate schools for black and white children.

To be sure, one can find some of this summary only be reading between the lines of the papers. But that, of course, is because these papers have other purposes and because of the monumental difficulty of dealing with such complex matters within so narrow and necessarily simplified a frame. By looking briefly at a very tiny piece of the whole, I would like to attempt to emphasize some of the legal ambiguity and the related interaction between law and education that is involved in the material considered in these papers. As a minute illustrative focus, I will take a test designed to determine a student's "readiness" for participation in a work-study program. Let us assume that 80 percent of the white students and 60 percent of the black students "pass" the test used; and let us assume a legal challenge based on discrimination against blacks.

If this legal challenge is founded on the equal protection clause of the fourteenth amendment of the United States Constitution, a Supreme Court decision (*Washington v. Davis*, 426 U.S. 229 (1976), poses a major obstacle. According to that case, the fact that a significactly higher proportion of black than white applicants fail an employment test does not, without more, show that the test was racially discriminatory; proof of a discriminatory purpose is required. The Court has also said, in *Washington v. Davis* and subsequently, that a challenger could prove the required discriminatory purpose indirectly. To this end, statistics showing a racially disproportionate impact would be relevant but not conclusive factual information. The Court noted in this respect the case of *Yick Wo v. Hopkins*, 118 U.S. 356 (1866), in which the disproportion was in the order of 99 percent to 1 percent. That reference, plus the fact that the disproportion in *Washington v. Davis* was 57 percent to 13 percent, suggests that the 80 percent to 60 percent imbalance of the illustration would provide relatively weak evidence of discriminatory purpose.

In *Debra P. v. Turlington*, 474 F. Supp. 244 (M.D. Fla. 1979), discussed in both papers, a federal district court discussed and ultimately distinguished *Washington v. Davis* in connection with its considerations of a challenge to Florida's competency test for high school graduation. The district court in *Debra P.* conceded that neither the disproportionate incidence of failure rates (ten blacks to one white in that case) nor the fact that the responsible education officials had anticipated this disproportion demonstrated a racially discriminatory purpose. But a distinction was found in the fact that the black students who were challenging the test were assumed to have suffered educational disadvantage attributable to school segregation. This critical fact provided the basis for a legal conclusion that the past discriminatory purpose to segregate schools was perpetuated by the present competency testing program. Yet, the locus of *Washington v. Davis* was in the District of Columbia, where the Jim Crow practice of separate but equal facilties for blacks and whites was prevalent in the public schools and other aspects of the city's public life. Since this background was not persuasive to the Supreme Court in deciding whether a discriminatory racial purpose was shown (or that its absence should be discounted), it is not obvious that the background of de jure school segregation should have been persuasive to the court in *Debra P.* Just as the inferior education of segregated schools might explain a lower rate of passing Florida's competency test, the Supreme Court has explicitly observed (in *Griggs v. Duke Power Co.*, 401 U.S. 424 (1971)) that such segregated education would disadvantage blacks in employment testing.

The *Washington v. Davis* precedent is significant because it makes proving the existence of a racial classification so difficult. That would not—technically—defeat the equal protection-based discrimination challenge to the hypothetical test which determines work studies eligibility. The challenger could in any event argue, correctly, that the test for admission to the work studies program is government action that classifies—between those who pass and those who do not—and that only government classifications which allocate benefits (or burdens) reasonably are consistent with equal protection. But, in the absence of a racial classification (or some other

ingredient which performs the same function and is assumed here not to be present), the reasonableness of the classification is made virtually invulnerable because of the controlling significance given to the institutional roles of court and educational tester.

With no proof of purposeful race discrimination (or the equivalent), the courts "defer" to the educational tester because the courts believe that our political system gives educators the role of making the critical judgment about what is reasonable. That is, in the court's view, the educators are empowered to judge the reasonableness of the classification resulting from the test, and the courts lack the competence as well as the authentic power to second-guess that judgment. As ordinarily framed, the governing legal principle requires the challenger to "prove" that there is no rational basis relating the classiciation (test passers vs. test failers) to the legitimate purpose of the test (e.g., to select those who would profit, or profit most, from the work-study program). It is generally conceded that the challenger will be able to meet the test so infrequently that the challenger's probability of success should be rated at 0.

Let us assume now that the challenger founds the challenge not on the constitution, but on some statutory and/or regulatory provisions that prohibit racial discrimination in providing work-study opportunities of the kind in question. Under this assumption, *Washington v. Davis* is not a direct barrier. Furthermore, the *Washington* opinion reaffirmed *Griggs v. Duke Power Co.*, which had held that, under Title VII of the Civil Rights Act of 1964, an employment test having a racially disproportionate impact is illegal if not validated. To validate a test the user must show that the test is an effective device for selecting the more qualified employees. In general, the different *Griggs/Washington* results are explainable in terms of differences in institutional roles and of the different implications of constitutional and statutory decisions. In *Griggs*, Congress had deliberately singled out employment discrimination based on race as an area of concern; in Washington, by contrast, the Court had no such policy decision to rely upon. Furthermore, the *Griggs* result—prohibiting unvalidated tests because of their disproportionate racial impact—was confined to the employment focus of the statute; by contrast, a disproportionate impact decision in *Washington* would have had sweeping implications over a broad range, including such far-reaching areas covered by the equal protection clause as criminal law, taxation, and welfare.

All of this suggests that the disproportionate impact challenge in our illustration might find smooth sailing if it is based on a statute or regulation. That conclusion is far from inevitable, however. Title VI of the Civil Rights Act of 1964 provides the most obvious statutory sources of such a challenge, but the rationale distinguishing *Griggs and Washington* may not favor the challenger relying on Title VI. That statute does not represent a deliberate policy judgment that racial discrimination in vocational education —or even in education generally—should be singled out as an area of concern; Title VI applies to all programs receiving federal financial assistance.

As a consequence, any adoption of a disproportionate impact principle for Title VI could have broad application over many areas. In fact, a majority of the Justices of the Supreme Court have indicted in *University of California Regents v. Bakke*, 438 U.S. 265 (1978), that the anti-discrimination principle in Title VI is identical to the anti-discrimination of the equal protection clause.

Reliance on the regulations issued under the Vocational Education Act would appear to face comparably difficult problems. The act itself contains no anti-discrimination provisions (based on race) and plainly does not represent a deliberate Congressional policy decision to prevent race discrimination in vocational education. The regulations under the Vocational Education Act do expressly prohibit racial discrimination, but these regulatory provisions draw their authority, not from the Vocational Education Act, but from Title VI. Unless there is some reason to read the

regulations more broadly than their authorizing legislation, it would seem that the anti-discrimination principle of *Washington v. Davis*, the equal protection clause, and Title VI would also apply to these Title VI based regulations. In fact, in the *Bakke* case, a majority of the Court evidently gave little significance to Title VI HEW regulations which tended to support racially conscious affirmative action to overcome the effects of past discrimination. And, in other recent decisions, the Supreme Court has not been willing to follow anti-discrimination regulations that seem to range beyond the scope of authorizing legislation. See *Southeastern Community College v. Davis*, 99 S.Ct. 2361 (1979) (handicapped); *General Electric Co. v. Gilbert*, 429 U.S. 125 (1976) (employment discrimination).

Let us assume, now, that the difficulties considered here can be overcome and that the racially disproportionate impact resulting from the work studies test of the illustration would require the test to be validated. At this juncture, a court would be faced with a second basic issue, and this issue combines legal and non-legal elements. The court must set itself up as something of an expert in testing. This might be accomplished in various ways—by the court's actually acquiring the expertise itself, by its use of a court-appointed expert, by relying upon the expert witnesses and/or arguments of the parties. But the court must accomplish this somehow. That is, somehow, the court must put itself in a position to understand what is meant by such things as construct validity, content validity, criterion-related validity; it must be able to decide which of these techniques is appropriate; and it must understand whether an appropriate technique has been correctly used. But it is not accurate to think of the court, simply, as assuming the role of a testing expert. In the end, a legal requirement of test validation poses a legal test. There is no automatic identity between something like "acceptable professional standards" and "acceptable legal standards"; the law may require more or less, or it may not. For example, it may be argued that even though the test in question meets professional standards, there is an alternative test (or an alternative to the test) that, at the same time, would be effective in selecting students ready to profit from work study, but would have a significantly lesser tendency to exclude black students. The court must decide whether such a less restrictive alternative test is legally required, and the court must decide what should be accepted as a sufficiently effective alternative selector and as having a sufficiently reduced racial impact.

Although it is accurate to characterize these decisions as ultimately "legal" decisions to be made by the court, it would certainly be misleading to imagine that the courts would ordinarily be free of the influence of the "real experts" in making those decisions. The extent of this influence on the court's decision defies prediction (and, perhaps, even defies accurate after-the-fact assessment).

Both the narrow illustration of this comment and the more far-ranging discussion by Tractenberg and Pullin lead to several clear implications for performance testing in vocational education. First, legal challenges to these programs will be made, both because perceived injustices are involved and because the legal machinery is at hand. Second, accurately predicting the outcome of the legal cases that will be brought is well beyond our collective wisdom at the present time. Third, the certainty of lawsuits and the uncertainty of results will feed upon themselves and create a distinct, though unknowable, reality of its own. This new creation will be shaped by two evolutionary processes—the one identified by the courts' episodic attempts to understand the world of education and testing and to articulate legal rules responding to that understanding; the other identified by the on-going attempts of educational planners to anticipate (and to avoid) the "worst" and of litigators to anticipate (and to exploit) the "best" of the emerging legal doctrine.

Part of our conventional wisdom, based on the insights of de Tocqueville, is that political questions sooner or later become judicial questions in the United States. What is seldom noticed is that, in the process of assimilation, the underlying issues are changed and distorted — initially in their new legal setting and eventually in themselves.

CHAPTER FIVE

IMPLEMENTATION ISSUES

The successful implementation of any program or product is not an easy task. Care must be taken that the steps in any implementation plan be carefuly identified and analyzed. These concerns are addressed in Chapter Five.

First, H. Brinton Milward discusses performance testing as an organizational innovation—not in the conventional sense of the term (i.e., measuring the performance of a student) but "rather of the performance of an entire training program of an instructor." He introduces such concepts as "ideas in good currency" as a necessary precondition to the adoption of an innovation. The remainder of the paper focuses on the diffusion and adoption of the innovation within an organization, the role of "street-level bureaucrats" in implementation and a technique—"mapping backwards"—to arrive at an estimate of what will be needed to successfully implement a program or practice.

In contrast to the organizational theory perspective presented in Milward's paper, Curtis R. Finch addresses the implementation from a more pragramatic point of view. He describes the implementation setting and identifies a series of considerations which should be kept in mind: curricular, teacher and ancillary personnel, administration, student and community. The points raised by both contributors are discussed by Janet E. Spirer in the Comments paper.

Performance Testing as an Organizational Innovation

H. Brinton Milward
University of Kentucky
Lexington, Kentucky

Performance tests are not innovations in vocational education. Vocational education has long given a broad variety of performance tests to certify students for particular occupations and trades. Thus, these tests are not an "innovation" in the conventional sense of the term that refers to a product or practice new to the adopting unit.

This paper will be concerned with the actual or potential use of the results of performance tests, not as measures of the performance of a student, but rather of the performance of an entire training program or of an instructor. From this perspective performance tests are an innovation that states, school districts or the federal government could use to evaluate the performance of instructors or programs. Thus, in the context of this paper, a performance test would not be an innovation to a welding teacher; it would be an innovation to the staff of the state office of vocational education who would use the aggregated results of students' scores to evaluate how successful a given program was in actually training people for specific occupations and trades.

The impetus for using performance tests as instruments of vocational education program evaluation comes from the implementation of the 1976 Amendments to the Vocational Education Act of 1968.[1] The act stipulates that both the Bureau of Occupational and Adult Education and the states shall audit and review vocational programs to make sure they are the best possible programs of vocational education. The role given to the states is very explicit:" . . . each State shall evaluate, by using data collected, wherever possible by statistically valid sampling techniques, each such program within the State which proports to impart entry level job skills. . ."[2]

In other words, vocational education must become result oriented. Increasingly, the emphasis will be on what the students can do in the occupations they have been trained for, rather than an evaluation that is oriented toward the process by which students have been trained.[3]

A response to the legislation and the general concern for government accountability has been improved monitoring of the effectiveness of training. Performance testing can be used as one mechanism assessing outcomes of the training process. In the past, performance tests have seldom been used in this fashion, and most evaluation efforts in vocational education have been". . . too casual, informal and fragmented and have only rarely served the cause of program improvement. . ."[4]

What is occurring in vocational education is no different from what is occurring in a variety of other programs. The federal government is attempting to increase the analytic capability of the states ". . . to strengthen state leadership in education, to put more of the monitoring

responsibility in the hands of state education agencies."[5] This has resulted from the federal government's inability to monitor or control effectively the behavior of thousands of programs scattered across the country. In intergovernmental relations, the federal government has become a provider of funds and a writer of guidelines and regulations. The states' role has become that of federal program managers, and the local programs are the delivery agents. This explains why the role of the states in evaluation of performance of local programs has become such a salient issue.

Ideas in Good Currency

There is a direct connection between adoption of an innovation and ideas in good currency. Ideas in good currency are a necessary precondition for the adoption of public policy innovations. The space race of the 1960s as well as the law and order movement of the late 1960s and early 1970s are examples of ideas of good currency. "Among their most characteristic features are these: they change over time; they obey a law of limited numbers and they lag behind changing events. . ."[6] The "failure of the schools" idea, which has led to accountability measures like minimum competency testing, is the idea in good currency behind performance testing as an instrument of evaluation.

New ideas in good currency usually emerge from a disruptive event in a series of events. These perceived crises set up a demand in society for new ideas to solve these problems. It is at this point that ideas which are beyond the mainstream of the public agenda begin to surface. This occurs through a process of diffusion that depends upon interpersonal networks and upon the communication media which in turn, shape the idea to their needs. In the use of minimum competency testing, traditionalists used "the failure of the schools" idea to try to abolish many of the non-traditional courses and programs which were developed in the 1960s and '70s.

Ideas must gain entry to the limited set of channels through which formal policy agendas are set. As Schon wrote, ". . .they require, in the approval of administrators, commissions, notable personages, legislators and the like, a kind of benediction."[7] This power is used sparingly and the decision to do so comes usually from a shared calculation of the idea's relation to personal and political interests and of the support the ideas have already gathered. As Feller, Menzel and Engel found in the case of federal legislation, the adoption of a new technology by a state was directly traceable to the passage of a new highway or air quality act. "Although federal legislation seldom mandates the adoption of a specific technology, the 'choice' may be narrowly defined."[8] Thus, ideas in good currency can affect diffusion patterns through an intergovernmental network.

Implementation as an Interorganizational Process

Innovations based on ideas in good currency must diffuse and be adopted, as well as implemented, into practice through an interorganizational network. There are two features of the vocational education network that distinguish it. First, since education is a state, rather than a federal function in the United States, there is no national configuration to the network in terms of equivalent organizations, actors or practices. Second, the delivery system for vocational education and training is difficult to distinguish from the general network of education.

For the most part, the vocational education delivery system is the same system used to educate all of the secondary, postsecondary, and adult students. School principals, superintendents, presidents, directors, and boards—those who make the decisions for education in general—also make the majority of the decisions for vocational education.[9]

There are a large number of organizations that shape policy and delivery of services in vocational education. These include, for example, the Department of Education, State Boards of Vocational Education, State Advisory Councils for Vocational Education, Local Boards of Education, the Department of Labor, State Employment Security Agencies, and CETA prime sponsors. In addition, these organizations exist in thousands of communities, fifty states and the territories and at the national level of government. Instead of a neatly arranged hierarchy with clear lines of authority, what we have is a loosely coupled functional system with considerably more power at the middle and bottom than at the top. In addition, all three levels of organization—federal, state and local—possess certain scarce resources valued by the others. Each level also has a certain amount of constitutional and behavioral independence.

While the organizations providing vocational education training are loosely coupled with those providing coordination and guidance, the network of actors and organizations consist of a tightly coupled policy network—albeit one which lacks elaborated, hierarchical authority relationships. It is a network which is boundary maintaining and which has persisted for over fifty years as a separate entity from the larger general education system. This occurred because of the differences in orientation, as well as in status, between the two groups. Vocational education is best described as a functional system that consists of:

1. The set of persons who lack, but want or need occupational skills or training.

2. The set of agencies, groups, and institutions that serve and train them.

3. The research, evaluation, and training activities that affect the provision of educational training.

4. The laws, policies, and programs under which vocational education is provided.

To call this a 'system' is not to imply that it has well-defined, consensual goals and coordinated programs for reaching them. The institutions included in [this] system tend, in fact, to behave in a fragmented and disorganized way."[10] Even though disorganized, this is the system with which persons seeking vocational training must deal.

Karl Weick calls this a "loosely coupled system where the individual organizations in the system are more like holding companies than goal directed entities." He suggests that this may be due to the diffuse task vocational education performs and the uncertainty of the technology used in the process of educating students.[11]

Performance Tests as an Evaluation Method

We are assumming here that performance tests are not an innovation to those who will administer them. Thus, a second assumption may be made; i.e., if the innovation is to be adopted and effectively used, then one must focus, not on the process of innovation diffusion, but rather on the implementation of the results of performance tests to local administrators, state vocational education officials and federal adminstrators of the Bureau of Occupational and Adult Education. With this as the focus, several corollaries must be spelled out.

First, any new system of evaluating programs or individuals will increase the programs' and individuals' uncertainty in regard to their performance. Uncertainty is a key concept in both organization theory as well as economics. A person or organization will always try to reduce the

amount of uncertainty they must deal with. People in public organizations, like vocational education programs, will prefer to be evaluated by instruments that they both understand and influence. Performance testing depends on measurable outcomes. Since it is the students' score that is aggregated, rather than superiors evaluating whether or not an instructor followed the correct process of teaching and test administration, the teacher may feel that the outcome measure of evaluation is unfair since a variety of things may affect the students' score. Too many students may be scheduled to take the tests at one time; the quality of equipment may vary from program to program; teachers may feel that they have more than their share of undermotivated students. Thus even if promotion, pay, and transfers are not tied to an evaluation system heavily relying on performance testing, it would be a major source of uncertainty for those being evaluated.

This suggests why an innovation model is not appropriate for performance testing as an evaluation instrument. Many of the innovation models assume that all innovations go through a sequence of stages approximating the research and development process where technology dominates the results.[12] This does not apply to educational innovations, like performance testing, as an evaluation instrument. The technical superiority of the innovation is very difficult to show and, in addition, the innovation clearly threatens both teachers and administrators whose programs will be evaluated with the information they provide. With educational innovations where the technology is "soft," implementation, not the superiority of the technology will dominate outcomes.[13] Education is not unique in being dominated by the implementation process.

Simply because teachers and administrators adopt an innovation does not mean that the adopted practice will be the same as the original innovation. The actual "outcome" of the adoption of performance testing will greatly depend on how teachers and administrators implement it. In a federal system, there are no command and control mechanisms for forcing compliance with directives from either the state or federal level. Interdependence and bargaining inevitably shape intergovernmental relations. In this case, as in so many others dealing with implementation, the "street-level bureaucrats"—the teachers—will largely determine whether or not the evaluation system produces meaningful information upon which to base program choices.

Street-Level Bureaucrats and Implementation

The concept of street-level bureaucrats is very important in understanding the introduction of an innovation into continuing practice. Street-level bureaucrats include teachers, police officers, welfare workers, public health officers, and many others. All of these officials work with the public and make decisions on the basis of individual initiative as well as established routine. They interact directly with citizens, in fact, they are most people's only direct contact with the government. Since they exercise considerable discretion in their jobs, they effectively determine how policy is delivered to citizens.

In other words: "To accomplish these required tasks, street-level bureaucrats must find ways to accommodate the demands upon them and confront the reality of resource limitations. They typically do this by routinizing procedures, modifying goals, rationing services, asserting priorities, and limiting or controlling clientele. In other words, they develop practices that permit them in some way to process the work they are required to do. [Their] work . . . is inherently discretionary. Moreover, it is difficult to establish or impose valid work-performance measures, and the consumers of services are relatively insignificant as a reference group. Thus street-level bureaucrats are constrained, but not directed, in their work."

These accomodations and coping mechanisms, that they are free to develop, form patterns of behavior which become the governmental program that is "delivered" to the public. In a significant sense, then, street-level bureaucrats are the policy-makers in their respective work arenas.[14]

As Weatherly and Lipsky and Richard Elmore point out,[15] this turns the study of both innovation diffusion and the process of implementation on its head. The lowest level of the implementation network determines policy while the upper-and mid levels are only able to circumscribe the behavior of lower officials within certain broad limits. This occurs because in a loosely coupled, interorganizational and intergovernmental network, goal homogeneity in the absence of hierarchical authority cannot be assumed "Interorganizational problems arise largely from the difficulty of coordinating the activities of several different units, each of which has its own goals and established routines."[16]

There appears to be an inverse relationship between the number of required transactions between organizations to implement a new program or practice and the likelihood of the implementation being successful. "Even when the probability of a favorable result is high at each step, the cumulative product of a large number of transactions is an extraordinarily low probability of success."[17] A recent study lays out in elaborate detail the multitude of devices and ploys that experienced administrators can use to subvert, deflect or delay the effect of programmatic innovations they do not like.[18]

Given the fact of the inability of state and federal officials to control the behavior of local teachers effectively, what can be done to increase the probability that an evaluation system at least partly based on performance tests will not be subverted? One technique for arriving at an estimate of what will be needed to implement a new program or practice successfully is called "mapping backwards." It proceeds from our assumption that power over the delivery of vocational education training effectively lies in the hands of the street-level bureaucrats—the teachers—rather than in the hands of administrative officials at higher levels of government. "In the bewildering variety of local institutions . . . one factor remains constant: The point at which public policy meets the private preferences and choices of young people is in individual contacts between teachers or program operators and young people. This is the street-level contact that determines whether policy affects the behavior of individual young people."[19]

Mapping backwards focuses not on the goals of the administrators at the top, who wish to use performance tests to determine which programs are successful and which ones are not; it begins with looking at the behavior of those who will be implementing the performance testing system then proceeds to ask the question "what do I want the teacher or local administrator to do?" Once that question is answered, one traces back though every step in the implementation process and at each step determines what needs to be done to increase the probability that a teacher will implement the performance testing system in the prescribed manner.

When the vocational education network is viewed from the bottom up, it becomes clear that whatever policy we wish to implement ultimately will depend, not on a centralized command and control system, but on changing the behavior of local teachers and program operators who actually deliver services to trainees.

The true policy problem that must be faced is not to make teachers behave consistently with respect to a new evaluation system, but to increase the probability that the teachers skill, judgment, and knowledge will affect the ability of trainees to find meaningful and productive work.

Conclusions

The preceeding sections described the network through which a performance test based evaluation system would be implemented. The paper has also identified where the ability to shape policy lies and whose behavior must be changed if a new evaluation system is to be successfully implemented. This section will focus on what can be learned about implementation and innovation from this discussion. The implication of the paper thus far is that ". . . the process of framing questions from the top begins with an understanding of what's important at the bottom."[20]

With the implementation of any innovation, there are three reasons to cooperate with those promoting the innovation. The first reason is self-interest. People and organizations join together because participants perceive the innovation to be in their best interest. Given the variety of different people and organizations in vocational education, it is unlikely that one innovation will be perceived in the interest of all or even a majority of the organizations and people in the network. Therefore, this is not a sufficient base on which to structure cooperation.

A second reason for cooperating is that higher level authorities mandate cooperation. Innovations that are linked to the governance system of an organization will obviously command more attention than those that are not. But a mandated evaluation system that has to be implemented across governmental boundaries and where the institutions involved are loosely joined will not have the same force as it would if it occurred within one organization.

A third reason for cooperation is exchange. Here, people cooperate because they receive something they value in exchange for their cooperation. In a loosely joined network this will facilitate cooperation, as it is unlikely that any one organization will have all of the resources needed to accomplish their tasks. This creates a positive incentive for mutual exchange of needed resources.

In reality, all three of the reasons for or inducements to, cooperate will be effective in certain situations. Also, the three are ideal types, and most interorganizational transactions have elements of more than one of the three; often one sees an organization adopt a "carrot-and-stick" approach to inducing cooperation.

The purpose of defining the three reasons for cooperation is that administrators at federal and state levels, when they are dealing with local officials, often assume that the local officials' interests and goals, are or should be, the same as their own. They also may operate as if an authority relationship existed between them and local officials. As this paper has pointed out, these are incorrect assumptions and may contribute to the failure of an innovation, such as an evaluation system to be implemented or, if implemented, to provide meaningful data on which to judge program performance.

If we wish to increase the probability of the implementation of a performance testing system as an evaluation instrument, we need to map backwards in our analysis from the teacher who will actually give the tests to the local administrator of the program, to the state vocational education officials in charge of evaluation, to the federal administrator in the Bureau of Occupational and Adult Education. This is the reverse of the process that most analysts propose. Systems analysis, policy analysis, and other rational techniques advocate starting with the goals of federal officials and mapping forward to the point of service delivery. In the absence of hierarchical control and common goals, this will not usually be effective.

If we map backward though, we find teachers who feel a great deal of uncertainty over a new method of evaluation that they cannot completely control. There are administrators of local programs and school principals who will wonder where the resources will come from to collect and tabulate the data generated by the system. These administrators will also know that teachers will put pressure on them to upgrade the equipment used for performance testing so it will be appropriate for the newly developed tests.

> "Any kind of broad mandate that occupational competence be demonstrated by vocational education students could be viewed as some kind of disaster. The reason . . . is quite simple: The mandates always seem to require more than can be produced under the constraints which exist."[21]

All of these pressures may dispose a local administrator to oppose or subvert the new evaluation system. With service delivery and people-processing programs you simply do not get implementation without resources. It is a necessary but not sufficient condition.[22] The sufficient condition is support for the innovation by the local administrator. In two different review articles on the implementation of innovations, one that specifically focused on the implementation of evaluation findings, the support of the local administrator was found to be critical in successful implementation.[23]

Given that vocational education is a bottom-heavy system, what suggestions can be offered to improve the chances of successful implementation?

1. Map the delivery network backwards from the activities of the teachers to the source of the innovation.

2. Often local administrators do not comply with a mandate because it is not accompanied by the resources to implement it. Try to distinguish between an unwillingness to comply and a lack of capacity to comply.

3. Only attempt to change those activities for which it is possible to specify a clear standard of performance.[24]

4. Attempt to intervene as closely as possible to the point of service delivery so that the innovation is not distorted in the levels between point of service delivery and the source of the innovation. There must be careful preparation of local personnel so that they are prepared to implement the new system. Their advice is also needed in shaping the innovation.

5. Rather than simply monitoring compliance, state vocational education agencies should emphasize services to local programs.[25]

6. While state and federal agencies cannot control the implementation process, they can differentially reward those local programs making the greatest effort to implement the innovation. The creation and manipulation of a program's incentive structure may be one of the more effective ways to increase the probability of successful implementation.

The central point admnistrators that should bear in mind is that while some policies, like affirmative action, are regulatory in intent, vocational education exists primarily to deliver services. Here compliance, while important, is secondary to improving the ability of schools and institutes to deliver services, the quality of which depends, to a great extent, on delegated control.[26]

Notes

[1]Public Law 94-482, October 12, 1976. 90 State. 2187.

[2]Ibid., Section 1[illegible](6)(1)(B).

[3]William W. Stevenson, "The Educational Amendments of 1976 and Their Implications for Vocational Education," Information Series No. 122. (Columbus: Center for Vocational Education, The Ohio State University, 1977), p. 5.

[4]Henry Borow, *Philosophical, Practical, and Technical Issues Pertaining to Performance Testing in Vocational Education*, unpublished manuscript, (Columbus: National Center for Research in Vocational Education, The Ohio State University, 1978), p. 7.

[5]Samuel Halperin, "Emerging Educational Issues in the Federal City," Occasional Paper No. 42, (Columbus: National Center for Research in Vocational Education, The Ohio State University, 1978), p. 7.

[6]Donald A. Schon, *Beyond the Stable State* (New York: Norton, 1971), pp. 123-124.

[7]Ibid., p. 140.

[8]I. Feller, D.C. Menzel, and A.J. Engel, *Diffusion of Technology in State Mission-Oriented Agencies* (Center for the Study of Science Policy Institute for Research on Human Resources: Pennsylvania State University, 1974), p. 31.

[9]The National Center for Research in Vocational Education, "The Status of Vocational Education: School Year 1975-76," Research and Development Series No. 162 (Columbus: Ohio State University, 1978), p. 64.

[10]Schon, *Beyond the Stable State*, p. 43.

[11]Karl E. Weick, "Educational Organizations as Loosely Coupled Systems," *Administrative Science Quarterly* 32, no. 2 (March, 1976): p. 1-19.

[12]For example, see Ronald G. Havelock, Planning for Innovation (Center for Research on Utilization of Scientific Knowledge, Institute for Social Research, University of Michigan, Ann Arbor, Michigan, 1973).

[13]Paul Berman, "The Study of Macro and Micro Implementation," *Public Policy* 26, no. 2 (Spring, 1978): p. 161.

[14]Richard Weatherly and Michael Lipsky, "Street-Level Bureaucrats and Institutional Innovation," *Harvard Educational Review* 47, no. 2 (May, 1977): p. 172.

[15]Ibid., p. 173 and Richard F. Elmore, "Mapping Backward," Paper presented at the annual meeting of the American Political Science Association, September, 1979, Washington, D.C.

[16]Robert S. Montjoy and Laurence J. O'Toole, Jr. "Toward a Theory of Policy Implementation," *Public Administration Review*, 39, no. 5 (September/October, 1979), p. 273.

[17]Elmore, "Mapping," p. 10. He is reporting the results of a study of the implementation process by Jeffrey Pressman and Aaron Wildavsky, *Implementation*, Berkeley: University of California Press, 1973), pp. 87-124.

[18]Eugene Bardach, *The Implementation Game* (Cambridge, Mass: MIT Press, 1977).

[19]Elmore, "Mapping," pp. 29-20.

[20]Ibid., p. 21.

[21]J. Stanley Ahmann, "Implications of the Minimum Competency Testing Movement for Performance Testing in Vocational Education," unpublished manuscript (Columbus, Ohio: National Center for Research in Vocational Education, 1979), p. 20.

[22]Weatherly, "Street-Level," p. 182.

[23]William L. Hull, "Implementing Evaluation Findings," Manuscript (Columbus: Center for Vocational and Technical Education, Ohio State University, 1970), p. 3 and Donald C. Orlich, "Federal Education Policy," *Educational Researcher* 8, no. 7 (July/August, 1979): p. 6.

[24]Richard F. Elmore, "Complexity and Control," Institute of Governmental Research, Public Policy Paper No. 11 (Seattle: University of Washington, 1979), p. 42.

[25]Weatherly, "Street-Level," p. 195.

[26]Elmore, "Complexity," p. 9.

Considerations in the Implementation of Performance Testing

Curtis R. Finch
Virginia Polytechnic Institute and State University
Blacksburg, Virginia

In our society, frequent change is inevitable. Employment opportunities shift, new occupations are established, and employers revise their expectations of workers. Change has also become quite prevalent in education. Rich,[1] for example, notes a variety of educational movements and innovations that have been proposed over the past two decades. Among these are the open classroom concept, career education, and mainstreaming.

In recent years, the notion of educational change has fallen into disrepute. This state of affairs is at least partialy due to teachers' perceptions of benefits derived from it. During the 1950s and 1960s, teachers were strongly encouraged to accept change and cooperate with others to ensure that it occurred. They were often told that a change would result in certain benefits such as greater efficiency or increased student learning. This, of course, did not occur in some cases, and teachers rapidly became disillusioned with change for the sake of change.

While a simple definition of change may be any alteration in the status quo, this does not take the basic concerns of educators into consideration. A more expansive definition must be used for educational change. It may thus be thought of as any significant alteration in the status quo that is intended to benefit the people involved.[2] Such a definition reflects the need to implement only those changes that have the greatest potential for positive payoff.

This paper examines one such change, giving consideration to its implementation in vocational education settings. Performance testing appears to have great potential for improving the educational process and the results of that process. However, its potential may never be realized if educators and others are not attentive to factors that hinder implementation in the schools.

As the other papers have noted, performance testing is a rather complex phenomenon. And once philosophical, legal, and technical issues surrounding performance testing have been at least partially resolved, there is still the need to deal with a host of implementation considerations. They include the basic implementation setting as well as the curriculum, teachers, support personnel, administration, students, and the community. Each of these areas will be examined in order to highlight some of the key issues associated with implementing performance testing in vocational education.

The Implementation Setting

When change is being considered, it may be most beneficial first to examine the setting in which change will take place. Hull, Kester and Martin[3] note the three elements that can provide

the necessary stimulation for change to occur. These include the change advocate, the targeted consumer, and the innovation. In the application of this basic notion to performance testing, consideration may also be given to several other key elements, namely the curriculum and the community.

The change advocate serves as an initiator of the change process. Logically, if change is to occur, some person, group, or organization must provide initial support. Vocational education administrators and supervisors tend to be most readily classed as change advocates; however, it is best to go beyond these individuals and consider others such as vocational teachers, ancillary personnel, students, parents, employers, and even professional organizations.

A second key element is the targeted consumer. Consumers are those who will actually use the innovation, not merely pass it on to others. They may, likewise, be persons, groups, or organizations. While the change advocate is hopeful that all consumers are eager to accept change, this is typically not the case. Some consumers are more adoption prone than others and are thus more receptive to change.

The innovation, which constitutes a third element, may have almost any form, dimension, or substance. In this instance, performance testing is reflective of a system that may be utilized as a basis for instructional improvement, evaluation, and accountability. If the Hull and Wells[4] scheme for classifying vocational education innovations were applied, it might be difficult to determine whether performance testing would be individual-behavioral, organizational-legislative, or scientific-technological. Classification may, in fact, be a function of the intended use and associated technology of performance testing.

Of equal relevance to change is the vocational education curriculum. Any educational change must be woven into the curriculum in such a manner that it is accepted and utilized. In terms of performance testing, thought should be given to a variety of areas including the alignment of tests, objectives, and the employment setting; varying technical content; and varying instructional settings. Each of these may affect the ways that performance testing is ultimately implemented in the schools.

The community is yet another element to be considered when change is taking place. Included in the community setting is a host of persons who must be dealt with various points in time. These include citizens, individual taxpayers, school board members, owners, managers, supervisors, personnel directors, and advisory committee members.[5] In this arena, concern tends to be expressed about the quantity and quality of education as well as how much vocational education will assist business and industry to grow and prosper. Community concern about change is extremely important since endorsement or lack thereof can spell success or failure. While individuals and groups in the community do not have day-to-day contact with vocational education, many are in a position to influence resource allocation and support for funding.

Curricular Considerations

The vocational education curriculum can be viewed as more than courses and content. Realistically, it reflects a broad range of educational activities and experiences. Given this perspective, we may define curriculum as "the sum of the learning activities and experiences that a student has under the auspices or direction of the school."[6] Thus included in the curriculum would be classroom, laboratory, and cooperative work experiences, cocurricular activities such as clubs and vocational student organizations, organized athletics, and music groups. It is within this setting that performance testing is intended to be implemented.

One basic curricular consideration has to do with the alignment of educational objectives, performance testing, and the employment setting. While educators have recognized for many years that instructional objectives for vocational education should be closely aligned with needs of business and industry, it has only been recently that organized groups have taken over the vocational teacher's responsibility to identify relevant objectives.

Consortia such as the Vocational and Technical Education Consortium of States (V-TECS) and the Interstate Distributive Education Curriculum Consortium (IDECC) have, in fact, worked toward the alignment of objectives and the work setting. This has consisted of developing objectives and (in the case of IDECC) learning activity packages (LAPs) that are based upon extensive task analyses and personal interviews with workers and employers. Given this situation, it appears quite easy to move toward performance test implementation (if it has not already taken place).

V-TECS, for example, has developed catalogs of objectives and criterion-referenced measures that might serve as a basis for test development. IDECC includes check sheets in many LAPs that can be used to evaluate student performance in applied settings. Of major concern is the potential that exists to develop tests that align with instruction, objectives, and job relevant content. The extent to which tests mesh with teacher and consortium efforts may well determine whether or not performance testing is accepted and used.

A second curricular consideration is that of test content variation. Performance test content varies as a function of curriculum content and, as such, may require different approaches to development and use. A close look at code numbers used for occupations in the *Dictionary of Occupational Titles*[7] reveals that workers have varying degrees of involvement with data, people, and things. For example, a salesperson would have a high degree of involvement with people, a computer programmer would work more extensively with data, and a welder would be more involved with things. While test developers tend to perceive such differences in tests, administrators and teachers may not be as aware of how curriculum content is translated into meaningful test content. If these variations are not taken into consideration, performance test relevancy may be seriously affected.

A somewhat similar situation exists with regard to the instructional environment. Tests and the testing process tend to vary as a function of the instructional setting. Thus, a test that is designed for use in a vocational laboratory may not be applicable to evaluation in cooperative employment settings. This could occur because students are paid for participating in a cooperative vocational program and report to an employer, whereas, in a school setting they are not paid and report to an instructor. In the school setting, instructors have complete control over the testing situation while in a cooperative setting this control is shared with employers. As the implementation of performance testing occurs, a close look needs to be taken at ways that tests can be adapted to different environments as well as what shared testing responsibilities may exist. This will at least partially alleviate some of the problems associated with testing in various instructional settings.

Mention must also be made of how performance testing may interface with the competency-based education (CBE) movement. While CBE has been in existence only a short time, its impact is being felt in all parts of the nation. Some states have, in fact, mandated the implementation of CBE by a specified date. Although CBE does not differ from other modes of education in terms of its goals, there are several key elements that serve to make it a powerful movement. These include using the competency (skill, attitude, value, or appreciation that is deemed critical to successful employment) as a basis for curriculum content, making available

explicit criteria for each competency, assessing competence in applied settings, having demonstrated competence serve as a determiner of student progress, and focusing on facilitation of student achievement of competencies.[5] It is clear that CBE and performance testing have the potential to work as a team, and in many functioning CBE programs that is the case. Any steps taken to implement performance testing should thus be coordinated with existing or proposed CBE activities. Obviously, it is much easier to effect one educational change than two separate changes!

Teacher and Ancillary Personnel Considerations

In many respects, teachers and ancillary personnel may be considered as the basic advocates and consumers of performance testing. These individuals are most likely to administer tests, determine results, and make professional decisions based on these results. Ancillary personnel include guidance counselors, placement officers, and similar specialists. These persons are in an excellent position to help students enroll in meaningful vocational programs and assist program graduates find employment. While teachers obviously have the major responsibility for performance testing in instructional settings, they are often heavily involved in student selection and placement activities and may work quite closely with ancillary personnel.

One basic consideration with regard to these groups is acceptance of the performance testing concept. Many may see performance testing as a threat to their positions; something that serves to hold them accountable for student achievement. Performance testing may be viewed by others as being no different from what is being done at the present time. This situation is particularly difficult to handle since professionals believe that they are already doing what is proposed. Others, however, might not be aware of performance testing's complexities and may only recognize their personal interpretations of the concept. Clearly, acceptance will be most difficult among persons who have misconceptions about performance testing. In fact, professionals who have had the least involvement with performance testing may be most eager and ready to implement it.

Running parallel to the acceptance concept is the expertise needed to conduct performance testing. Sanders[9] notes several potential problems associated with performance testing administration. These include control over the testing environment and standardization of testing conditions and scoring procedures. Test administration processes are reasonably common knowledge to measurement specialists and those who have had experience developing and administering valid and reliable performance tests. Vocational teachers, on the other hand, have not always been exposed to the psychometric properties of performance tests and how these properties may be altered through test administration. If performance testing is to be implemented in vocational education, the knowledge gap must be narrowed.

While teachers are not expected to become measurement specialists, they should at least have a working knowledge of factors that can affect test validity and reliability. A poorly developed and administered test is worse than no test at all. Consequently, any implementation scheme must deal directly with improving teacher knowledge and showing how this knowledge may be applied to realistic educational settings and testing situations.

Since some teachers and support personnel must be convinced to accept performance testing and to learn about its unique character, how may this task be accomplished? One logical approach consists of inservice education. Credit or noncredit workshops could be offered that provide educators with an awareness of performance testing, an understanding of its strengths and limitations, and an opportunity to conduct tests under the supervision of workshop leaders.

One key aspect of the inservice education process is motivation. If educators are not positively motivated to participate in inservice education, any proposed implementation may be doomed to failure. Palmer notes that both extrinsic and intrinsic motivation are used to encourage educators to improve their performance. With regard to extrinsic motivation:

> The impetus may come from rule enforcement (making participation in inservice programs a requirement of the job), or from rewards that are valued by the participants but do not stem from improved performance (such as bonuses, increments, certificates, etc.)[10]

Persons who related most closely to extrinsic motivation are those who have not yet satisfied their basis needs or do not obtain satisfaction of higher order needs from their work. As far as intrinsic motivation is concerned, "the impetus for improvement may come from a desire to do a better job of teaching. Intrinsically motivated teachers derive satisfaction directly for the performance of their teaching duties."[11]

Clearly, it would be desired that educators involved in performance testing inservice programs be intrinsically motivated. Some educators, of course, will not be motivated in this way and thus must be reached through extrinsic means. Then, once involved in an inservice program, these persons may become intrinsically motivated to implement performance testing in their vocational programs.

Administration Considerations

Even though teachers and ancillary personnel accept performance testing as a worthwhile concept and have been trained to use tests, the implementation process is by no means complete. There are several factors in the administration of a performance testing program that must be examined very closely. These factors can serve either to enhance or hinder implementation depending upon how they are handled. Among the more critical factors are testing scheduling, test facilities, determining students' grades, and communicating test results.

It is reasonably easy to schedule a classroom pencil-and-paper test. In this instance, students are all brought into the classroom, sit at different desks, and are each given a written test to complete. Performance testing takes on a somewhat different air. Students typically take performance tests individually or by small groups in laboratory or work settings. In most cases, actual equipment, materials, and people are used to make the test as realistic as possible. These requirements often place a heavy burden on vocational educators since it may be difficult to arrange test schedules in an acceptable manner and have adequate supervision available. In military technical training, where performance testing has been used successfully for over thirty years, scheduling is of major importance.[12] In fact, blocks of time for performance testing are built directly into students' training schedules, and instructors are assigned to coordinate and monitor testing activities. Time made available for testing may be as much as six hours and student-instructor ratios of six to one are typical. Given this situation, it is easy to see why performance testing in military settings is so successful. Students may be tested individually under controlled conditions under the watchful eyes of skilled instructors. They are placed in controlled environments before and after completing the test so that answers are not passed on to others.

The military testing model indicates some of the major scheduling problems that may occur when performance testing is carried on in school settings. While recognizing that military and

civilian vocational education do differ, educators should be aware of various scheduling concerns. The successful implementation of performance testing will require that answers be sought to questions such as. How many blocks of time should be scheduled exclusively for performance testing? What must be done to ensure a reasonably low student-instructor ratio during the testing period? How will test security be controlled before, during, and after students are tested?

As an alternative to scheduling blocks of time, teachers might choose to test on an individual basis whenever students appear ready. This may be quite easy to accomplish, especially when the performance is of a manipulative nature. As with group testing situations, it is essential for the teacher to use standardized equipment, materials, directions, and conditions. Additionally, in the case of tests that focus on fault diagnosis (e.g., electronic and automotive troubleshooting), a large number of representative troubles must be at hand. Otherwise, hints may be passed on from one student to another with the result being invalid test results.

A natural outgrowth of scheduling processes is the establishment of testing facilities. Numerous authors have emphasized the need for a facility or area that may be used exclusively for performance testing. Wilson indicates that it is "highly desirable if a regularly assigned space could be set aside for conducting performance tests."[13] Performance testing facilities help to ensure that uniform conditions are set up for all examinees. While this notion may seem farfetched to vocational educators in the public schools, it is one which should be seriously considered. Having uniform test conditions allows all examinees an equal opportunity to do their best work. Within the testing area, it is extremely important to have equipment and materials which are the same from one test administration to another. If examinees are tested with non-equivalent equipment and materials under varying conditions, test scores will not reflect performance against a standard criterion. Vocational educators must recognize the need for standardization in testing processes and adhere to these standards whenever tests are administered to students.

Although not always associated with performance testing, the issue of grading is often raised when student achievement is to be measured. While most teachers would agree that grades serve few useful purposes, grading is an integral part of our educational process and as such, must be dealt with as performance testing is implemented. Of practical consideration is the way or ways that performance test results can be translated into a locally established grading scheme.

Teachers and administrators must reach some basic agreement as to how performance test scores will align with present grading policy or serve to modify that policy. This is not something that can be accomplished by an external advisor. Teachers need to consider, for example, what weighting may be applied to various tests and how this weighting contributes to determination of a final grade. Administrators must set up a system that ensures that students are being given appropriate credit for performance test completion. Other concerns will surely arise since local situations may point to a host of potential grading problems.

A final administrative consideration has to do with articulation. Performance testing has great potential to enhance communication between secondary and postsecondary institutions in terms of offerings, credit granting, and content. In fact, a properly administered testing program may enable students to receive advanced placement at community colleges and technical institutes. The articulation process (groups of persons from different institutions working together to ensure a minimum of course duplication and a maximum of transfer credit) seems very much in line with performance testing concepts. Tests can serve as communication devices that assist groups of educators to note exactly what is expected of students in various

educational settings. Thus, as performance test implementation takes place, a look should be taken beyond individual courses and schools to see how processes might be articulated with schools and programs at other levels (e.g., secondary, postsecondary, adult, CFTA)

Student Considerations

While students' needs and interests are often considered as vocational curriculum content is being established and teaching/learning strategies are being selected, this is generally not the case when tests are being devised. Apparently, some teachers have felt that testing is a secret process that must not be revealed to anyone until some appropriate time. Students are required to develop high levels of anxiety and engage in testing activities that are very unfamiliar to them. Obviously, if such practices are followed with regard to performance testing, the end result will be even greater anxiety and frustration. An alternatve to the possibility of utter chaos is placing greater emphasis on students' concerns and being sure that these concerns are built into the testing process.

Initially, it might be best to examine students' acceptance of the performance testing concept. Since some students have only taken pencil-and-paper tests, they may not understand what peformance testing is. For these students, it would be necessary to design some sort of orientation program that clarifies performance-testing procedures, provides each person with "hands-on" experiences, and generally relieves anxiety. This approach should serve to improve students' acceptance of performance testing and speed the implementation process.

A second consideration has to do with student contributions to testing. Students can be given opportunities to help design tests. For example, if a test involves "cutting a piece of metal with an oxy-acetylene-cutting torch," students might talk to welders about the standards tradespersons would use to evaluate such a cut. They might read technical manuals to determine meaningful process and product criteria. The information could then serve as a basis for evaluating student performance. Students would, thus, be more aware of how they are expected to perform and where test standards come from. Even though students are seldom involved in test design, the nature of most performance tests makes this procedure reasonably easy to carry out. It should not detract from the validity of most tests and will certainly reduce student anxiety.

A final student consideration has to do with evaluation of testing. All too often, teachers do not report to students about how well they perform on tests. Students do not like this sort of treatment, and it will affect their attitudes to any type of testing, including performance testing. While written nearly thirty years ago, Michaels' and Karnes' comments about performance testing are still very appropriate: "After the test has been administered and scored, discuss with the class outstanding strengths and weaknesses noted. Give the students an opportunity to ask questions and clear up any misunderstandings."[14]

Reporting results to students helps them understand the importance of time, efficiency, proficiency, quality, and similar performance criteria. It also serves to reinforce the importance of doing one's best work on a test and amplifies the need to follow test directions and procedures.

The Community

Even though individuals in the community may have little involvement with performance testing, they must not be left out of the implementation process. In this case, the approach taken is more akin to public relations with key groups in the community being informed about performance testing. With parents communication needs to be started early in the implementation process, and they should be told why performance testing is being used as well as what it means to their children. When a youngster comes home one day and complains loudly about a "weird" performance test, the parent should already have some notion about such tests. Keeping parents informed serves to strengthen support for performance testing in the schools, especially if those parents take an active part in reinforcing comments made to students by their teachers.

Most vocational programs enlist the assistance of advisory committees composed of business and industry representatives. These committees advise and assist vocational educators by verifying the need for instruction, examining course content, providing teachers with technical assistance, and providing various services to students, the school, and the community.[15] Any performance testing implementation plan should give consideration to these committees. This may range from informing members about performance testing to soliciting ideas for test development.

Advisory committees help to link education and work and, as such, can provide invaluable services. The vocational teacher should, therefore, draw heavily upon this resource whenever tests are being developed and revised. Assistance might consist of identifying appropriate work samples, identifying potential criteria, selecting equipment and materials, and reviewing testing and scoring procedures. Extensive involvement by advisory committees will contribute greatly to the solidification of community support since members tend to be key leaders in their respective occupational areas. Their support of the performance testing concept will be looked upon by other employers as a very positive sign.

Employers, other than advisory committee members, also need to be informed about performance testing. As the consumers of vocational education products (graduates), employers should have a basic understanding about how vocational students are tested and how test performance aligns with work performance.

In order to keep employers informed, some vocational programs have developed performance-based transcripts that indicate what the individual student is able to do in terms of tasks and skills rather than merely using a statement of grades. This approach lets the employer know what to expect of a program graduate and helps in determining the initial duties that persons will have on the job. The basic focus of performance tests can easily serve as a foundation for transcripts. Details such as the level of acceptable behavior and conditions might also be included for each listed item.

Employers appear eager to find out more about what potential employees can do, and performance testing has the potential to meet their needs, particularly if a meaningful communication device such as a performance-based transcript is developed and used.

Summary

Implementing performance testing in vocational education settings is a complex process Those responsible for implementation must take a host of factors into account and work with numerous groups and individuals if any sort of success is expected to occur. The character of vocational education demands that linkages be developed with persons in education as well as in the community at large. Teachers, support personnel, administrators, and students each have a role in performance test implementation. Failure to include one or more of these groups in implementation plans will most certainly work against the movement.

Finally, parents, advisory committee members, and employers play an important part in the implementation process. Their collective support ensures that performance testing will be recognized as being beneficial to persons outside of education.

The message is clear that implementing performance testing in vocational education will be a difficult, time-consuming task. However, given the many benefits derived from performance testing, any time devoted to implementation will be well spent.

Notes

[1]John M. Rich, *Innovations in Education, Reformers and Their Critics* (Boston: Allyn and Bacon, 1978).

[2]Ronald C. Havelock, *The Change Agents' Guide to Innovation in Education* (Englewood Cliffs, N.J.: Educational Technology Publications, 1973).

[3]L. Hull, J. Kester, and B. Martin, *A Conceptual Framework for the Diffusion of Innovations* (Columbus, Ohio: Center for Vocational and Technical Education, The Ohio State University, 1973).

[4]William L. Hull and Randall L. Wells, *The Classification and Evaluation of Innovations for Vocational and Technical Education* (Columbus, Ohio: Center for Vocational and Technical Education, The Ohio State University, 1972).

[5]Curtis R. Finch and Robert L. McGough, *Administering and Supervising Occupational Education* (Englewood Cliffs, N.J.: Prentice-Hall, in process).

[6]Curtis R. Finch and John R. Crunkilton, *Curriculum Development in Vocational and Technical Education* (Boston: Allyn and Bacon, 1979), p. 7

[7]U.S. Department of Labor, *Dictionary of Occupational Titles*, 3d ed. (Washington, D.C.: U.S. Government Printing Office, 1965).

[8]Finch and Crunkilton, *Curriculum Development in*, pp. 220-21.

[9]James R. Sanders, "Measurement Problems and Issues Related to Applied Performance Testing" (Paper presented at ERA, April, 1976.

[10]Teresa M. Palmer, "In-service Education: Intrinsic Versus Extrinsic Motivation," In Louis J. Rubin, ed., *The In-Service Education of Teachers* (Boston: Allyn and Bacon, 1978), pp. 215-19.

[11]Ibid., pp. 215-19.

[12]Robert Glaser and David J. Clouse, "Proficiency Measurement," In Robert M. Gange, ed., *Psychological Principles of System Development* (New York: Holt, Rinehart, & Winston, 1962), pp. 419-74.

[13]Clark L. Wilson et al., *A Manual for Use in the Preparation and Administration of Practical Performance Tests* (Washington, D.C.: Office of Naval Research, 1971), p. 48.

[14]William J. Michaels and M. Ray Karnes, *Measuring Educational Achievement*, (New York: McGraw Hill, 1950), p. 366.

[15]Center for Vocational Education, *Organize an Occupational Advisory Committee*, Module A-4 (Athens, Ga: AAVIM, 1978), pp. 8-9.

Comments on Implementation Issues In Vocational Education

Janet E. Spirer
National Center for Research
in Vocational Education
Columbus, Ohio

"The best laid plans. . ." is a well worn phrase that we have all heard and probably find ourselves muttering from time to time. It certainly may be applied to publicly funded programs where it is often acknowledged that there is a gap between policy intentions and policy implementation. Recognizing the tendency for this gap to exist—and often expand—is crucial, regardless of the policy or program being implemented. The two implementation papers present some concerns with which administrators and teachers must deal when implementing performance testing.

The authors broach the implementation issue from two different perspectives which appear to be complementary. Milward discusses the process by which an evaluation system, partially or completely relying on performance testing, can be implemented. He explains how ideas or issues come to the fore (i.e., ideas in good currency) and who should be involved in designing implementation strategies ("street level bureaucrats"). The major strength of Milward's paper lies in its generalizability. That is, administrators could apply the concepts Milward introduces to any program planned or currently in operation.

If an administrator sat down and as Milward suggests, "mapped backwards" to identify those persons who should be involved in the implementation process, the "considerations" addressed by Finch certainly would emerge. Finch's paper is written more pragmatically and should help an administrator begin to identify specific audiences (and what he terms "considerations") that might affect the implementation process. These include: curricular considerations, teacher and ancillary personnel considerations; administration considerations; student considerations; and community considerations.

Thus, while Milward's paper introduces the process by which an administrator implements any evaluation system, Finch provides the reader with a "laundry list" of who and/or what "considerations" might affect the implementation of performance testing. However, a note of caution is appropriate. While the implementation process is generic, each vocational education program or school exists in an individualized environment with its own set of actors, constraints and problems. Therefore, Finch's "considerations" should serve only as the first step when "mapping backward." This handbook, as a whole, deals with other considerations that might prove to be as, if not in some cases, more important for a specific vocational education program or school. For example, some legal considerations, especially if a state has adopted a minimum competency testing law, might be crucial to successful implementation. Or, the institution of performance tests that are not proven to be valid and reliable might undermine the entire implementation process.

Also, the purpose behind performance testing--an evaluative tool to improve programs and student learning--should be focused on as the implementation process is designed and then carried out. Dissatisfaction with evaluation's usefulness has produced an extensive body of literature contending that evaluation seldom influences program decision-making. However, studies have been reported that deviate from this stream of thought. For, example, Michael Q. Patton, Edward C. Weeks, and Marvin C. Alkin, et al[1] have made strong cases for the usefulness of evaluation by adopting a broader definition of utilization.

The literature is replete with suggestions for increasing the utilization of evaluation information. For example, Weeks[2] offers three factors thought to influence the use of evaluation findings: (1) organizational location, (2) methodological practices, and (3) decision context. Alkin, et al[3] have identified eight factors affecting the utilization of evaluation information. These include: (1) preexisting evaluation bounds, (2) orientation of the users, (3) evaluator's approach, (4) evaluator credibility, (5) organizational factors, (6) extraorganizational factors, (7) information content and reporting, and (8) administration style.

Regardless of whether one subscribes to Weeks' model, Alkin et al's model or other models appearing in the literature, inherent in all of these models are factors which need to be carefully identified and defined in order to implement a performance testing program. Milward offers "mapping backward" as a method to identify the concerns and their interrelationships. Finch's "considerations" often will surface in this process. However, the point to be made here is that no author can identify, *a priori*, the actual considerations that will be appropriate in every setting. These papers describe the implementation process and some considerations that may be appropriate. But the final list of considerations that emerge when the implementation process is conceptualized and then carried out must be individualized to meet the specific needs of a vocational education, policy, program or school.

Notes

[1]For example, see Michael Q. Patton, *Utilization-Focused Evaluation* (Beverly Hills, California: Sage Publications, Inc. 1978); Edward C. Weeks, "The Managerial Use of Evaluation Findings," in H.C. Schulberg and J. M. Jerrell(Ed) *The Evaluator and Management* (Beverly Hills, California: Sage Publications, Inc., 1979) pp. 137-255; Maurice C. Alkins, Richard Daillak and Peter White, *Using Evaluations* (Beverly Hills, California: Sage Publications Inc., 1979).

[2]Weeks, "The Managerials Use," p. 139.

[3]Alkin, et al., *Using Evaluations*, p. 235.

CHAPTER SIX

IMPLICATIONS FOR VOCATIONAL EDUCATION

The first paper by Robert E. Spillman and Charles D. Wade begins by exploring different perceptions of vocational education (e.g., human resources view, humanistic view, social reform view and general education view). They then discuss why four issues—legal mandates, human resource needs, student needs and institutional and curriculum concerns—are important for vocational education. The paper concludes by offering the response they feel vocational education must make to the philosophical, technical, legal, and implementation issues raised in the handbook.

In the second paper, Nellie Carr Thorogood also deals with the question of implications for vocational education. Using a different approach from Spillman and Wade, she looks at the role of "stakeholders" in vocational education and performance testing, the uses of performance testing in vocational education and discusses the implications of the issues raised by the contributors by delineating those internal to and external to the institutions. A third perspective on the implications of the four issues for vocational eduction is presented by Marvin R. Rasmussen in the Comments paper.

The Implications of the Issues for Vocational Education: A Viewpoint

Robert E. Spillman
Charles D. Wade
Bureau of Vocational Education
Frankfort, Kentucky

Introduction

Performance testing is a tool which can be used by vocational educators to improve the quality of programs, enhance the learning process by students, and strengthen the accountability of vocational education. However, it is not without problems or limitations, but with careful planning the process can be effectively implemented into vocational education programs.

The purpose of this chapter is to review the major issues in performance testing identified by the authors of the previous chapters and to bring into sharper focus the implications for vocational education.

The contributors to this publication agree with Slater's definition that "performance tests refer to tests in which the test stimulus, the desired response, and the surrounding conditions approximate the reality of an actual situation drawn from a specific occupational or role-based context."[1] Several of the contributors discuss in detail the variety of reasons for performance testing. The consensus seems to be an agreement with Slater's four major purposes: (1) formative program evaluation, (2) summative program evaluation, (3) instructional management and decision-making, and (4) student certification.

At this point, the reader begins to identify some conflicts among the philosophical, technical, legal, and implementation issues surrounding performance testing. To relate both commonalities and differences of the issues of performance testing to vocational education, some understanding of the purpose of vocational education is necessary.

Exploring Different Perceptions of Vocational Education

There is no widely accepted statement describing the purpose of vocational education. Although various documents from the federal government, state education agencies, and local institutions address the purposes of vocational education, no effort is made in this chapter to persuade the reader to accept or reject these purposes. Rather, this chapter will simply explore some different perceptions of vocational education.

Human Resources View. Some believe vocational education is responsible for supplying a pool of well trained people from which business and industry can select employees. This view requires that the graduates have entry-level job skills and appropriate attitudes that make them productive on the job and contributors to the economic growth of the community, state, and nation. In this perspective, service to the economic system dominates service to the individual.

Humanistic View. From this view, vocational educators are responsible for preparing all vocational students for employment in their chosen vocations. The needs and desires of the students are given major consideration in all aspects of the program. Students are challenged to achieve to the highest level of their potential, regardless of the local availability of jobs. From this point of view, the graduate, in a mobile society, seeks employment in a broader area and becomes a contributing member by being trained for maximum contribution. Curriculum decisions are more sensitive to individual needs than to local job market requirements.

Social Reform View. Recent federal legislation has highlighted this view by giving less attention to human needs and desires and more attention to increasing the enrollment of both sexes in nontraditional classes. Again, education is asked to be the leader in removing social deficiencies, such as discrimination based on sex, race, economic deprivation, and physical or mental handicaps. In attempting to meet these needs, vocational educators are often faced with conflicts when the community expresses resistance to the social reforms. Parents may not want their children in nontraditional programs, and employers may be slow to employ graduates for nontraditional jobs. The social reform approach maximizes access to all programs for any student and pressures traditionalists to accept contemporary societal goals

General Education View. This view acknowledges the need for the institution to assist students in making meaningful career choices; it also promotes the idea that specific job skills should not be taught in the institutional setting. In this view, the students should be given economic awareness, self-awareness, and career awareness, with the specific skill training left to the employer. Supporters of this concept believe all students should receive some orientation to a variety of occupations without spending extensive periods of time in developing competencies in a specific occupation. More time is spent socializing the students to the labor force than developing skills.

All of this leads up to the fact that the implications of performance testing for vocational education depend, not only on an understanding of performance testing but also on a perception of the purposes of vocational education. In Chapter Two, Borow discusses some fo the conflict that occurs between the goals of optimum human utilization and the objectives of maximizing personal potential.

Important Issues for Vocational Education

The intent of this handbook is to identify issues underlying performance testing as they relate to vocational education. Perhaps one question which should be asked is why vocational educators are concerned with performance testing at this time. In Chapter Five, Milward clearly states that performance testing per se is not an innovation in vocational education. The brief history of performance testing in the Preface indicates that this form of testing has been acknowledged and, in fact, used by vocational educators for many years. The answer to the current concern may be found in the new degree of sophistication in the tests, testing procedures, and test analysis and in the innovative uses of performance testing. Why these issues are important for vocational education can be discussed in four areas: (1) legal mandates, (2) human resources needs, (3) student needs, and (4) institutional and curriculum concerns.

Legal Mandates. While Public Law 94-482—the Vocational Amendments of 1976—and its resulting regulations do not specifically require performance testing, it is certainly a method to

be considered in addressing the requirements for program evaluation. Section 104.402 of the Rules and Regulations states:

> "The State Board shall, during the five year period of the state plan, evaluate in quantitative terms the effectiveness of each formally organized program or project supported by Federal, state, and local funds. These evaluations shall be in terms of: . . .
>
> (b) Results of student achievement as measured, for example by:
>
> (1) Standard occupational proficiency measures;
>
> (2) Criterion referenced tests; and
>
> (3) Other examinations of students' skills, knowledge, attitudes, and readiness for entering employment successfully."[2]

State boards have struggled with this area of evaluation. Performance testing has not been widely accepted as a program evaluation tool. Slater's summative program evaluation description is appropriate for describing the utilization of performance testing for program evaluation. As indicated by Milward, performance testing for program evaluation is innovative and must encounter the implementation problems that he and Finch address in Chapter Five. According to Pullin, there may also be legal implications, such as a situation in which program quality requires termination of an instructor's contract.

In three-fourths of the states, legislatures have considered some form of minimum competency testing, according to Tractenberg. A few states, by policy and regulation, have mandated competency-based vocational education and its related curriculum-based performance testing. Borow describes a relationship between competency-based programs and performance testing. As these programs grow in acceptance, states are mandating local participation.

Student certification in occupations seem to be increasing. Performance testing for student certification in vocational areas has generally been limited to the health and personal services areas such as nursing, cosmetology, and barbering; however, licensing requirements for aviation mechanics and communication electronic operators have existed for years. Newer efforts include certificaton of fire fighters, emergency medical technicians, and automobile mechanics.

According to Pullin and Tractenberg, the area of student certification—and its legal implications—is a major concern. For those adhering to the human resources perception of vocational education, student certification is a positive step for any occupation, since it gives the employer greater assurance of hiring a quality employee. Persons with other views of vocational education may resist performance testing for student certification; however, new occupations may mandate such student certification for graduates who wish to work in those occupations.

Whatever one's perception of the purpose of vocational education, the legal mandates by the federal government, state governments, and occupational boards and agencies make performance testing a concern for vocational educators.

Human Resources Needs. For a large number of vocational educators, advisory committees, and business and industry representatives, needs for human resources deserve special attention.

If vocational programs are to be accountable to employers, students must be trained in the entry-level skills required for the job. In Chapter Three, Klein presents a model for determining job competencies as well as for developing performance tests. Proper performance tests can measure each student's job competency and the entire program's proficiency in relating to actual job requirements.

Not only do graduates need initial job skills, but they must also possess that difficult-to-measure trait called "employability." Borow discusses the need to include the affective domain in performance tests since many jobs depend on such things as attitudes and ethics; however, Tractenberg cautions that there are legal problems relating to the students' right to privacy when attitudes are included in the test items.

The first objective of vocational education graduates is to be employed, but they soon wish to advance to positions requiring greater skills, better human relations, and leadership ability. While the earlier writers do not stress need for leadership development, Borow states that "performance tests should be chosen and administered to measure competencies related to the aims of broad, liberal education as well as those of work."[3]

Employers apparently want workers with skills, but in line with the "general education view" of vocational education, they also want employees with job adaptability and advancement capabilities. Performance tests strive to simulate the actual job situation, but final evaluation may have to come with follow-up studies of both the employers and the graduates who have been placed on the job.

Student Needs. To vocational educators, social service agency personnel and advocacy groups of various types, vocational education can be the answer to the employment problems of most people. However, the goals of serving industry and meeting the needs of students are often in conflict. For instance, Borow notes the conflict between an open admissions policy and the use of certifying examinations. An open admissions policy is "humanistic," while student certification supports a "human resources" view. In addition, Pullin and Tractenberg agree there are problems associated with performance tests for student certification; i.e., in establishing performance standards, educators must maintain integrity with employers and, at the same time be aware of the possibility of discrimination to the student because of socioeconomic background, race, or sex.

Performance tests must be constructed to protect the rights of all students. Those who view vocational education as a "social reform" program see this as a major issue. In no case should performance tests discriminate on the basis of race, sex, handicap, or membership in a special population. Pullin and Tractenberg point out that using "instructional management and decision-making" for evaluation presents problems since the remedial program indicated by the diagnostic test could segregate the groups by sex, race, or type of handicap. Performance tests for summative evaluation can present a problem when classes or institutions have a disproportionate enrollment of special populations. The expectations for successful program completions may have to be altered when a large number of students are academically, mentally, or physically handicapped.

Institutional and Curriculum Concerns. Administrators of vocational programs must be concerned about the use of performance tests in their institutions. A good deal of controversy surrounds the uses of performance tests and who makes the decisions regarding their use. Performance tests may be good, but Borow raises the question, "for whose good?"

Teachers may not object to formative program evaluation when the purpose is to make program adjustments and curriculum improvement. Students may not object to "instructional management and decision-making" evaluation as long as it is used for prescriptive programming for instruction, but summative program evaluation affects the teacher personally, if the results indicate program termination. Student certification is also viewed with alarm by students who have spent up to two years in a program and then are rejected from the occupation by a final performance test. These kinds of serious concerns require resolution.

Institutional administrators must also be concerned about the cost of performance testing and the time allotted to testing. Finch stresses the need for performance testing to become a part of the instructional program with time blocks, space, equipment, and personnel assigned to this task. The military has used this approach for years and assumes it to be an important function of the instructional process. The competency-based vocational education movement incorporates performance testing concepts in the instructional program, since each competency must be mastered to the desired standard before the student can be recognized as having completed the task. Administrators and instructors must clearly identify the relationship between the competency-based vocational education curriculum and performance testing.

Response of Vocational Education to the Issues

In this section, the authors deal with the response that they feel vocational education must make to the philosophical, technical, legal, and implementation issues associated with performance testing. The topic is dealt with in six major subdivisions: (1) philosophical adoption of the concept, (2) test development and administration, (3) uses of performance tests, (4) access and equity, (5) curriculum improvement, and (6) implementation of performance testing.

Philosophical Adoption of the Concept

The fact that Willers and Borow did not quite reach agreement on a philosophical base for performance testing points out the need for each vocational education agency to proclaim its own philosophy of education formally before initiating performance testing. To be successful in this endeavor, educational leaders must develop general goals of education—including vocational education. These goals need not be measurable; in fact, the major purpose should be to set a direction for the organization that is consistent with its basic philosophy. Only those institutions that believe in job training should attempt to develop performance objectives for vocational education. Vocational educators should develop specific, measurable course objectives that are based on actual job needs and on well-established general goals.

While some narrowly define performance tests as measures of psychomotor skills only, developers and users of such tests would be well advised to include cognitive competencies and, when the technology permits, the affective domain. It should be noted that the regulations for P.L. 94-482 indicate a need to measure "students' skills, knowledge, attitudes, and readiness for entering employment successfully." This challenges educators to develop measures to address the "whole person." When performance tests do not measure the cognitive and affective domains adequately, vocational educators should supplement the test with other methods of evaluating these domains.

There is no merit in having a "pure" performance testing system if it does not meet the needs of the student and the institution. State and local vocational agencies should supplement

performance testing by developing and implementing an extensive follow-up system. Such a system should determine the extent to which vocational graduates are placed in the occupations for which they are trained. The follow-up should also assess the extent to which employers are satisfied with the training received by their employees. Analysis of such data should be useful in supplementing performance testing. Since future funding may be contingent on how well graduates perform in the actual occupation, this type of data could prove to be invaluable.

Attention should by now have been directed to one major reason why performance testing should be adopted—and while many good reasons may be discussed, one top priority must be the desire to achieve accountability. Accountability is the dominating force in modern decision-making at the policy, legislative, and budgetary levels. Regardless of which of the views of vocational education are held by educators (most probably accept a combination of all four), vocational education does deal with selecting, preparing for, and securing a job. Vocational education assists people in moving from a life focused around school to a life focused around a job. It serves to bridge the gap between school and work for many people. To this end, accountability deals with the extent to which the program assists students, through successful employment, to become contributors in the economic system.

Agencies and institutions that recognize the basis for performance testing and are willing to supplement testing with other appropriate measures should find testing beneficial in documenting the accountability of vocational programs to the public and to the policy makers.

Test Development and Administration. Vocational education must respond to the technical aspects of performance testing by developing acceptable measurement instruments and administering these tests in a manner that stands scrutiny by professionals in the testing field. The performance tests must meet the tests of validity and reliability.

Klein and Perloff discuss the relative difficulty of developing performance tests. Vocational education performance tests should be based on actual occupational needs and be representative of on-the-job situations. In this regard, much work has already been done that should ease the developmental process. The Vocational-Technical Education Consortium of States (V-TECS) has developed many catalogs of performance objectives through a rigid research process that ensures that the most important tasks performed by workers are included. If both the curriculum and performance tests were developed using an approach similar to that of V-TECS, the effectiveness of the developmental process, as well as its cost, should be more pleasing to administrators.

Performance tests may vary in their degree of sampling but the critical aspect should be predictability of the test. A variety of testing approaches, such as direct work observation, work sample, and simulation should be used to ensure that the performance tests assist educators in viewing the students as they should function in the actual job setting. Tests should be criterion referenced in order to measure the level of competence against the standards of the occupation.

Uses of Performance Tests. Each segment of the vocational education community must carefully study Slater's purposes of performance testing and identify those areas that will be most important in its program. For example, performance tests given before student enrollment in a program may be used for screening or diagnostic purposes. However, screening will be permitted in only a very few programs operated by public educational institutions. The legal issues noted by Pullin and Tractenberg can generally be avoided if the tests are used for diagnostic purposes, in order to prescribe a meaningful instructional program for each student.

In addition to performance testing before student enrollment, tests can also be very valuable during the course of student programs. For instance, during a program, performance tests can be used effectively for both student and program diagnostic purposes. In-route testing of skills should reduce the likelihood that students could spend months in a program only to learn near the end of their program that they are unable to pass the performance tests. Also, with student diagnostic tests, provisions can be made for remedial programs and services early in the program. Performance tests for program evaluation purposes should direct teachers and administrators to make program adjustments without long delays.

Finally, administration of performance tests at or near the end of the program permits both the certification of students and summative evaluation of programs. In the future, there may be more occupations for which licensing tests are mandated. In the meantime, vocational educators can use performance tests as a means of describing the tasks that students can perform. The test score may not always be used to determine successful completion of a program; rather the score can describe students' skills when they leave the programs. The end test can also be used to make program changes and, in some cases, terminate programs not meeting standards.

Vocational educators, educational planners, and legislative bodies must use care in analyzing the results of performance tests. Test data can be very useful in improving vocational programs; however, the tendency must be resisted to misuse the data in ways such as limiting enrollment of those predicted to fail by the performance test or terminating programs based solely on test performance of the graduates. Care must also be taken not to misuse the concepts of performance testing; i.e., abusing the rights of students and teachers by expecting more from the results than the test is capable of giving.

Access And Equity. The problems of access and equity are often created by inappropriate and unrelated criteria for entrance or acceptance in a program or a job. Sex or race are not appropriate criteria for assessing ability to do a particular job. The concepts of performance testing should provide an opportunity to overcome many of the issues of access and equity. Properly validated performance testing—not race, sex, socioeconomic background or other discriminatory criteria—should measure ability to perform the job. Graduates of vocational programs who possess certification that they possess the competency necessary for a particular job, have a valuable bargaining tool in seeking employment. Certification provides an opportunity to focus the employment interview on documented competence, rather than on social bias.

The concepts associated with validation of performance testing must provide assurance that there is a direct correlation between the content of the instructional program and the content of the test. Whether students are admitted to or complete the program should be based upon their ability to perform identified tasks and not upon other unrelated criteria.

If it is used properly, the performance test will enhance education rather than victimize students and instructors. Proper use can be accomplished by adhering to the guidelines for fundamental fairness, due process, and [illegible]cy as described by Pullin and Tractenberg. Statewide standards, established by a recognized governmental agency, administered responsibly, and used properly, should promote access and equity in vocational education.

Curriculum Improvement. The greatest value in performance testing may be its potential for improving the instructional programs. Competency-based vocational education programs are based on the same job analysis concept as performance testing. Rather than simply sampling job skills, the competency-based curriculum requires that students be tested on objectives for all job skills associated with their program of study. The catalogs of performance objectives from V-TECS can be used to produce competency-based curricular materials and performance tests.

With the development of performance testing, many educational institutions now recognize and grant credits for competencies that students have acquired outside the institutional setting Learning does not begin and end with formal schooling in an institution. The need to address this, as far as credentials are concerned, has been a recent development. It may be, in part, a response by educational institutions to the problems of declining enrollment, to the desire of many adults to return to school for more formal education, and the need to articulate programs between levels of education. At any rate, performance testing provides an opportunity for vocational educators to serve the needs of students and employers better as well as add efficiency to the vocational education system With well-validated test items, students may skip parts of the instruction in areas in which they have developed competence from other experiences. Education interrupted by personal situations or family needs may be resumed without loss of time and resources.

The development of performance testing may lead to performance contracting to provide vocational education services. Private industry can identify specific groups of people who need specific competencies. Contracts can be negotiated with educational institutions to provide these services with the understanding that if the students do not perform, the budget will be reduced accordingly. By using these concepts, vocational education programs may assist governmental agencies seeking to solve problems such as youth and minority unemployment and training for displaced homemakers.

This type of "individualization" of the curriculum to fit the needs of students can also be achieved by fitting the instruction to the learning rate and style of the individual student. Performance testing can allow the students to progress at their own rates and the instructors to select teaching strategies best suited to the needs of each individual student. In addition, performance testing provides the instructor, as well as program evaluators, some means of assessing the extent to which each student achieves the desired goal (employability) regardless of the route taken to that end.

Implementation and Performance Testing. Vocational educators tend to do things in a systematic, orderly manner and consequently, usually have much success in implementing new programs. However, the implementation strategies suggested by Milward and Finch should even further improve the possibilities of successful implementation of a new concept in an existing program. Implementing performance testing will be easier if Milward's "street level bureaucrats" are in support of the concept. To involve the teacher in the basic inservice program will meet the criteria of intervening as closely as possible to the level of the delivery system. Total involvement of students, parents, faculty, administrators, and the community at large is most desirable. Perhaps the local administrator, more than any other person, has the greatest influence on successful implementation of any educational concept. The administrator can assist staff members to do backward mapping in planning for implementation.

At the state level, vocational education must respond by providing leadership in implementation—including enthusiastic promotion, inservice training of staff, and most importantly, assurance that adequate funding is available from some source. Mandatory requirements for performance testing should be avoided and some differential reward or some other palatable means should be used to secure local cooperation in implementation.

Conclusion

Obviously, there are many issues to be considered concerning the why, who, and when of performance testing in vocational education. Certain philosophical, technical, legal, and implementation issues remain to be answered if performance testing is to be useful and effective as a professional tool to enhance the teaching/learning process.

The vocational community of administrators, teachers, counselors, teacher educators, curriculum specialists, and others must respond to the challenge as they have on so many other occasions. While some, no doubt, will reject performance testing altogether, others will find its appropriate use in their own vocational educational agencies and institutions.

Notes

[1]Stephen J. Slater, "Performance Testing: An Overview."

[2]U.S. Department of Health, Education and Welfare, Office of Education. Federal Register. Vol. 42, No. 191. Washington, D.C.: U.S. Government Printing Office, 1977.

[3]Henry Borow, "Performance Testing and Social Responsibility: An Issues Analysis."

Implications of Performance Testing on Vocational Education

Nellie Carr Thorogood
San Antonio College
San Antonio, Texas

Performance testing has been defined in this handbook as an applied testing process that is designed to measure performance on tasks requiring the application of learning in an actual or simulated setting (see Slater's discussion in Chapter One). Vocational education performance testing has chiefly been defined as a measure of competency in some specified field of occupational or career training, according to Borow in Chapter Two.

In a period of history in which economic impact and development is of major concern to the nation at large, education is being asked to provide more experiences related to the workplace. Richard Bolles indicates that "work and education alike have as their common task the business of teaching, refining, and using skills and knowledges."[1] Perhaps more than ever, there is increasing demand for vocational education to be more responsible for this economic development by providing greater reality to the workplace, and facilitating education to individuals. Performance testing is a clear route to the measurement of outcomes to be achieved by vocational education students, instructors, and programs. However, the use of performance testing in vocational education is not without implications and concerns. This paper will attempt to review the issues and the major implications for the utilization of performance testing in vocational education.

Stakeholders in Occupational Education and Performance Testing

In his book *People at Work*, Pehr G. Gyllenhammar introduced the term stakeholders to refer to persons or groups who have a "stake" or "interest" in the achievements and well-being of the company. He wrote:

> "The company must administer the resources with which it is entrusted . . to create economic growth, taking into consideration all the interest groups involved with the company. This includes consideration not only of the stockholders and the managers, but the customers, the supplier, the employees, the government, and society as a whole."[2]

Stakeholders in vocational education could include students, taxpayers, practitioners (teachers, administrators, counselors), state governments, federal government, employing institutions and the community at large. The issue papers presented within this handbook indicate these stakeholders in the vocational education program will be involved in the performance testing process. Involvement of stakeholders in the implementation of performance testing indicates the need for practitioners to consider the following types of activities:

- A clearly defined plan for the use of performance testing within the vocational education program.

- Articulation of the goals of this process among the key groups involved – students, secondary and postsecondary schools, businesses, industries, governmental agencies, and communities at large
- Clarification of the relationship of the ongoing programs to the new process. How does performance testing relate to the existing program? What will be the use of performance testing in admissions to programs, progression through the programs, and graduation?
- Active solicitation of involvement and commitment from the stakeholders in the new process
- A continuous flow of information to those concerned with the utilization of performance testing.

The primary stakeholder affected by performance testing in vocational education is the student. The implications of performance testing for the students are that the skills, knowledges, and competencies intended to be mastered can be measured and verified via performance testing.

Performance testing involves the student in an active role within the measurement process—the student is asked to perform, to show mastery of skills and knowledges.

In the book, *Carl Rogers On Personal Power*,[3] several trends are identified that appear to be occurring:

- Toward the exploration of self, and the development of the richness of the total, individual, responsible human.
- Toward the prizing of individuals for what they are, regardless of sex, race, status, or age.
- Toward human-sized groupings in our communities, our educational facilities, our productive units.
- Toward a more genuine and caring concern for those who need help.
- Toward creativity of all sorts—in thinking and exploring.

These represent exciting trends, ones that are appropriate to vocational instructors and their students. These trends represent the need for the human being to be literate, to be functional, to be productive, and to integrate into the environment in which he or she lives. Performance testing can be a positive accountability process for students while they are in a vocational education program, but more importantly it can be a valuable process to use throughout life in assessing one's ability to perform. Much of any occupational task is performance and most of us are completing a performance test daily.

Abraham Maslow[4] described a series of assumptions concerning human beings in his book of notes entitled *Eupsychian Management*. Some of these assumptions are of importance to the practitioner—both instructional and administrative—who will implement performance testing:

- Assume everyone is to be trusted.

- Assume everyone is to be informed as completely as possible.
- Assume in all employees and students the impulse to achieve
- Assume that people are improvable.
- Assume that people prefer to feel important, needed, useful, successful, proud, respected, rather than unimportant, interchangeable, anonymous, wasted, unused, expendable, disrespected.
- Assume a tendency to improve things; to make things better, to do things better.
- Assume performance for being a whole person and not a part, not a thing or an implement, or tool, or "hand".
- Assume the preference for working rather than being idle.
- Assume all human beings prefer meaningful work rather than meaningless work.
- Assume the preference for personhood, uniqueness as a person, identity.

Utilizing these assumptions places all of the stakeholders in an active, constructive, participative role rather than a passive, accepting, or destructive role. The student is actively involved in mastering the skills, knowledges, and competencies. The practitioner is actively involved in linking the student with the occupational setting through appropriate and meaningful instruction. Finally, the publics are actively involved in the input to instructional processes as well as in employment of the students.

The intentional outcome of performance testing can be:

- Improved student skills, knowledges and abilities
- Improved measuring and accountability processes for occupational instruction
- Improved productivity at the occupational job site.

The by-product of performance testing is the focusing, by all stakeholders on improved human competence.

On Competence

The overriding implication from the issues presented in this handbook is the idea that vocational educators who use performance testing will have to continually focus on quality, quantity (productivity), and costs of this process. Quality will have to be concerned with accuracy, as well as accomplishment beyond mere accuracy such as market value, quality judgment points, physical measurements and quality of "worklife" ratings. Quantity will need to include the rate of productivity, the timeliness of the criteria utilized, the appropriateness of ways utilized, and the volume of the "how many" question. Cost factors will include human resources

for research, development, implementation, and revision; materials involved in research, development, implementation, and revision; and the management involved in supervision, information flow, and assesesment of the process

No matter what the issues concerning performance testing in vocational education programs, the focus will need to continue to be *competence*—competence of knowledge, competence of skills, and competence of applications. Thomas Gilbert defines human competence as a function of worthy performance.[5] If vocational education leads to competence and competence is linked to performance, then at some point in time vocational education must be concerned with the assessment of performance.

Uses of Performance Testing in Vocational Education

If vocational education is to provide students increased opportunities for employability, three critical uses can be made of performance testing—advisement, instructional monitoring and assessment, and certification of competencies. These uses of performance testing can occur in classrooms, laboratory settings, simulations, or at the workplace.

It is important to keep in mind that performance testing is but one part of the advisement process; is but one part of the instructional process; and is but one part of the certification process. However, it can provide the basis for the planning of the entire vocational instructional process. The general goal of vocational education is access to employability. The general goal of performance testing in vocational education can be to provide clearer advisement; clearer feedback and direction in instruction; and more realistic certification of competencies to facilitate access to employability.

Kenneth Hoyt defines employability to include the following skills, knowledges, and abilities:[6]

1. the basic academic skills of mathematics, oral and written communications
2. good work habits leading to productivity in the workplace
3. a personally meaningful set of work values
4. a basic understanding of the American economic system
5. an understanding of one's own vocational interests, aptitudes, and abilities as well as opportunities
6. skills needed to choose a career
7. job-seeking, job-getting, and job-holding abilities
8. discovering unpaid work as a productive way to spend leisure time
9. capacity to affect positive changes in occupational society
10. skills needed to humanize the workplace and move up an occupational ladder

The implications of utilizing performance testing as an ends unto itself is a narrow approach and would have significant legal implications for practitioners. Performance testing needs to be one of the valuable tools of process in focusing occupational education on human competence. Performance testing must be part of an instructional process that includes clear identification of intended outcomes; utilization of appropriate materials, strategies, and experiences to facilitate the intended outcomes; and application of appropriate procedures and instruments to assess and measure the student progress (performance testing can be one of the most appropriate procedures). Once the student has completed this instructional process, the individual, the instructor, and potential employer will have clear information concerning the skills, knowledges, and abilities that have been achieved.

The challenge for vocational education programs within this decade appears to be to maximize the resources available in order to provide the best quality of programs to a diverse clientele. The programs will have to be flexible to meet the diversity of student needs. Many innovations, accountability structures regulations, and guidelines have been suggested in order to facilitate the vocational educator's ability to produce this maximization.

However, one of the educator's overriding needs to meet this diversity and challenge will be improved information. Improved information has the potential for creating greater competence in the day-to-day implementation of vocational education. The process and product of performance testing can be one vehicle to improve such an information flow.

Review and Implication of Issues

The implications of the issues presented in these papers can be reviewed by identifying the issues that are internal and external to the institutions that provide vocational education. Given the definition of performance testing presented by Slater, the following factors are important to issues that are internal to the implementation process: organization type, technology, purpose of testing, task to be accomplished, and organizational resources. In addition to these factors, there are also factors external to the implementing organization (environmental factors) that will have significant implications for the implementation of performance testing in vocational education. The external factors include technical, political, economic, legal, social, cultural, historical, and philosophical arenas. These internal and external factors will interact and impact the implementation of performance testing.

Internal Factors Affecting Implementation of Performance Testing

Concerning the identification of the implementation factors internal to the organization, the issue papers indicate the following:

- The *purpose* of performance testing for vocational education. This is the central and most critical factor. The purpose of the utilziation must be identified and clearly defined for the implementing organization. The purpose needs to be clearly articulated.

- The *tasks* that are involved in the implementation process to fulfill the goals and the purpose.

- The *practitioners* who will perform the tasks.

- The *resources* that will be needed to perform the tasks--included are human resources, physical resources, and fiscal resources. Resources include those that are internal to the organization and those that may come from business, industry, and other areas of expertise

- The *technology* that is necessary to perform the tasks. Included is the scientific content, the methodology content, and the process.

- The *organization type* and *structure*. The organization type can be a local school or training center, a school or college district, a state agency, a federal agency, or professional association. Organization structure will include all of the factors that are considered within the implementing structure--authority, decision levels, and so forth.

It is important to note that all of the factors internal to the organization depend upon a clear identification of the purpose of performance testing in vocational education. Once the purpose is clearly identified, then the practitioners are responsible for implementing the tasks with the highest amount of technology within the constraints of the organization's type, structure, and resources.

External Factors Affecting Implementation of Performance Testing

The presented topics have dealt primarily with the philosophical, technical, social, and legal issues confronting the implementation of performance testing within vocational education. There are additional issues in the implementation of performance testing including economic, political, cultural, and historical forces and factors. The latter will be defined briefly, but need to be considered in detail for future study.

Historical Forces and Factors. As with the utilization of any major technology and phenomenon, the historical elements are to be valued. Major historical issues impinging on vocational education and performance testing include the following:

- The traditional ways of preparing for work, that is, (1) organized apprenticeship—either voluntary or involuntary; (2) family teaching of a trade or craft;[7] and (3) the pick-up method by observation or imitation.

- The concept of the educated worker—both in the area of liberal arts and in occupational learning—has been a theme of vocational education since the beginning of the 20th century.

- The concept of performance as a measure of work productivity.

- Federal and state legislation.

- Technological developments.

- Knowledge development concerning: (1) the ways in which people learn, and (2) methodology of diagnosing learning needs and learning occurence.

- Vocational education's intention to relate to people and the work they do.

- The belief in the reality of individual differences in personal competencies and in the ability to observe them (see Borow, Chapter Two).

- Applications of institutional testing during the early part of the twentieth century including: (1) intelligence testing of children, (2) employment testing industry, (3) objective testing in the schools (see Borow, Chapter Two).

- Continuous use of performance testing in the U.S. military.

- Performance testing as the oldest form of evaluation of individual achievement (see Klein, Chapter Three).

Political Forces and Factors. Implementators of performance testing must always be aware of the political implications—power, control, "ownership" of standards, attitudes of major groups with a vested interest, opinion, and reactions to implementation tasks and technology.

Cultural Forces and Factors. Practitioners will need to consider cultural norms, cultural values, work place values and ethics, subcultures within society, public attitudes, social and cultural groups practices, and so forth.

Economic Forces and Factors. Whether the setting is a public or private institution, the general economics of the implementation process and tasks will need to be considered. The implementors must also be aware of the well-being of the general economy. For example, if additional financial resources will be required by an institution to implement performance testing, where will the funds be generated? what is the general economic indicator of the time? what is the unemployment rate?

All of these forces and factors need to be studied in greater depth for their implications for performance testing in vocational education. However, some of the most important factors and issues are found within the philosophical, technical, social, and legal arenas. The contributions and constraints to the implementation of performance testing from these areas have maximum implications.

Philosophical Forces and Factors. The values, ideals, ethics, and concepts that exist both internally and externally to vocational education will have direct implications on the use of performance testing. An exceptionally critical impact will be in policy-making at all levels and specifically within policy-making concerning the definition of the purpose of performance testing in vocational education. Philosophical issues include the following:

- Ideals of the models of performance testing models to be utilized.

- Integrity of the measures of competence.

- Commitment to the purpose and to the technical methods.

- Concepts and endeavors focused on the total well-being of the individual student.

- Conceptual purpose of performance testing to the whole of vocational education processes.

- Performance testing interface with ideals of the society such as democratic ideals, national priorities, welfare of the individual, worth of education, mission of schools, and open admission policy to institutions and programs.

- Concentration on outcomes of students, personnel, and programs.

Technical Forces and Factors (applicable to both the internal and external factors). Major technical issues of performance testing that can impinge on vocational education include

- The process for identification of competencies.
- The setting of standards.
- Objectivity.
- Validity-criterion, content, construct, consistency.
- Reliabilty.
- Application of performance testing (diagnostic; advisement; assessment; evalation; or certification).
- Utilization procedures (purpose, policy, and operational).
- Costs (dollar resources, human resources, time, physical resources).
- Quality of the competencies established and standards set.

Other technical issues include:

- The need for the performance tests to closely duplicate reality.
- The need for the skills, knowledges, and competencies required in an occupational field to be identified by persons in the field.
- The need for realistic, supportive test-related materials—instructional experiences and materials; laboratory experiences; simulations; work experiences.
- Observer and rater variability.
- Standardization.
- Efficiency of process, products, and procedures.
- Currency of tests in relationship to reliability and validity development process.
- Security of tests.

The actual construction of a performance test is a sophisticated and critical process. The steps offered by Klein are worth reviewing because the thoroughness aspect of the test development process has major implications for vocational education. The technology of this process will impact on all stakeholders in vocational education. It is important to consider this process as both dynamic and continuous if the performance testing used in vocational education is to be realistic to the workplace.

Social Forces and Factors. A variety of social issues can have an impact on the use of performance testing in vocational education. Most of these issues focus on the welfare of the individuals or groups and the ideals of the society. These issues include:

- Diversity in the needs of populations to be served-age, sex, ethnicity, learning abilities and disabilities, and various gradations of economic status.
- A move to look beyond just the needs of high school age youth to determine what is expected of vocational education.
- Ability to cope with nontraditional students.
- Technological displacement of employed persons.
- Learning experiences that occur as part of the normal process of work, community service, and life.
- Economic development and maintenance of communities in specific and of the country in general.
- Expected linkages between education and the place of work.
- Expectations of testing purpose (formative and summative program evaluation, instructional management, programing, and decision-making; and student diagnosis, advisement, achievement, and certification).

Legal Forces and Factors. Legal forces that surround performance testing in vocational education include the legal framework of the school; the local, state, and federal laws; decisions of the courts and quasi judicial bodies; and decisions and standards of regulatory agencies. In Chapter Four Tractenberg identifies seven legal concerns related to performance testing. They stem from:

- Federal and state due process.
- Federal and state equal protection clauses.
- Federal and state clauses protecting privacy and freedom of belief.
- State education clauses.
- Statutory laws.
- Regulatory laws.
- Common law.

Pullin identifies these major legal issues that are of concern in performance testing to include: student, personnel, and program accountability, due process in the use of performance testing, discrimination in the use of tests, and the right to privacy.

Implementation Forces and Factors. As Milward and Finch reveal (see Chapter Five) the implementation process is a complex one. Therefore, it is important to review from the papers some of the key implementation issues that will impinge on the use of performance testing in vocational education. These implementation issues include:

- Overall policy guiding the implementation process. What are the goals of intended outcomes? From what level are these goals generated—local, state, federal, other?
- Time involved in the development, implementation, evaluation, and revisions of the proceses and procedures.
- Costs of resources necessary for effective implementation.
- Quality of competencies, standards, tests, and utilization techniques.
- Quantity of the competencies, standards, tests, and utilization techniques.
- Methodology utilized—feedback, guidance, complementary education and training, reinforcement and remedial instruction, and assessing.
- The implementation setting including curriculum, teachers, support personnel, administration, students, employers and the community at large.

Each of these issues must be considered in relationship to the students, the practitioners, and publics who will be involved in the process and procedures.

Conclusion

In a period of time when lifelong learning, continuous development, career education, high-level technology, accountability, and emerging occupations are more than just sets of words linked together, the challenge for the utilization of performance testing within vocational education is critical. Since performance testing is not new to vocational education, the true challenge is to adapt performance testing to the diversities and demands currently being placed on vocational education programs. In meeting the challenge of these demands, performance testing may be used to assess prior learning and work experience so that the student can begin at the most appropriate educational level. Performance testing will probably continue to be used for certification in certain professions. Performance testing may be used for effective articulation from secondary to postsecondary levels. And, performance testing may be a vehicle of learning that is most closely related to the work situation. Afterall, productivity in professions, in businesses, in the trades, and in life generally is attuned to performance.

Therefore, vocational education programs through (1) a clearly defined plan of implementation; (2) a clearly defined plan of development and utilization of criteria; (3) a clearly articulated flow of information persons directly (students, practitioners, employers) and indirectly (taxpayers, governmental agencies, and citizens) involved with the process; and (4) continuous feedback system of information can effectively utilize performance testing as a *product* as a *process* of learning to achieve competence.

Performance testing is not a perfected process at this point in time. The potential use of performance testing in vocational education will depend upon the direction of the future of the

work place; the direction of education processes; and critically, the direction of technological advancements both high and appropriate technology. However, it has enough merits to be continued, to be improved, and to be utilized as a transitional process until more appropriate processes are developed. It is important not to let legal, social, cultural, and political constraints hamper the use of performance testing in vocational education. Historically, the purpose of vocational education has been educating an individual for gainful employment. A major vehicle utilized to produce these skills and competencies was performance—the ability to show in the laboratory or on the job an ability to produce and perform with competence. The implications of the issues presented in these papers indicate that performance testing will continue to be a vital alternative for vocational education. However, the practitioner of the future will be challenged to be clear in the definition of competencies, knowledgeable and sophisticated in testing methodology, and articulate in communicating all of the above to the stakeholders who have interest in vocational education in general and in performance testing specifically.

Notes

[1]Richard Bolles, "Training for Transition," *Education and Work* (Change Magazine Press, 1979).

[2]Pehr G. Gyllenhammar, *People at Work* (Reading, Massachusetts: Addison-Wesley Publishing Company, 1977), p. v.

[3]Carl Rogers, *Carl Rogers on Personal Power* (New York: Delacorte Press, 1977), p. 282.

[4]Abraham H. Maslow, *Eupsychian Management: A Journal* (Hometown, Illinois: Richard D. Irwin, Inc., 1965), pp. 17-36.

[5]Thomas F. Gilbert, *Human Competence: Engineering Worthy Competence* (New York: McGraw-Hill Book Company, 1978), p. 18.

[6]Kenneth B. Hoyt, "Employability: Are the Schools Responsible," *New Directions for Education and Work 1* (Spring, 1978): 30.

[7]Melvin L. Barlow, "Implications From the History of Vocational Education", (Columbus, Ohio: Center for Vocational Education, The Ohio State University, 1976), pp. 1-2.

IMPLICATIONS FOR VOCATIONAL EDUCATION: A THIRD POINT OF VIEW

Marvin R. Rasmussen
Portland Public Schools
Portland, Oregon

The purpose of the two papers in this summary chapter was to review the issues underlying performance testing and vocational education identified in the preceding chapters and to bring into sharper focus the implications of these issues for vocational education. This was all a task, and the two papers succeeded in accomplishing it to a varying degree. In the course of these efforts they have provided valuable additional perspectives on many of the issues addressed in the earlier chapters.

Thorogood's paper addresses some of the relevant issues. Others are unfortunately omitted or given scant attention. Early in her paper she acknowledges the relationship between competency-based education and performance testing. Both movements stem from the same social and educational sources—loss of public confidence in education and recognition of the special needs of the less academically talented students. Moreover, the two concepts are logically linked in that the "life skills" outcomes sought in competency-based programs often lend themselves well to performance testing and perhaps only to this form of measurement.

Two other related points that deserved more attention are: (1) performance testing needs to be integrated into the instructional process, and (2) performance tests cost more than conventional paper-and-pencil tests.

The crucial issues of the greater costs of performance tests as compared to standardized tests is only hinted at. It would have been useful to point out that the performance test is a more direct measure of student achievement and this tends to increase its validity and therefore its usefulness. But, this increase is purchased only at a substantial increase in the cost of testing in dollars and time. Great care needs to be used in deciding whether there is a real increase in validity and, if so, whether it is worth the increased cost over less direct but perhaps adequate measures.

Thorogood's discussion of the legal issues in performance testing identifies the major areas of legal concern and makes some useful suggestions for fairness and privacy. In reviewing Tractenberg's paper, Thorogood notes the legal implications of the key technical issues in performance testing.

Overall, Thorogood's paper was an incomplete but valuable contribution to the discussion of the issues surrounding performance testing and vocational education. Spillman and Wade's paper is comprehensive and insightful. Their valuable contributions would have been more accessible, however, if they had organized their discussion of issues by the categories used in the preceding chapters of this handbook. Thus, we would have had a discussion of each of the

philosophical, technical, legal and implementation issues and their implications, capped off by a summary of the implications for vocational education as well as for education as a whole. Instead, the section titled, "Important Issues for Vocational Education" is divided into "Legal Mandates," "Human Resources Needs," "Student Needs," and "Instructional and Curricular Concerns." While most of the major issues presented in the preceding chapters are touched on in the course of these discussions, it is difficult to hold them in perspective because of the organization of this section.

The subsection on "Legal Mandates" illustrates the organizational problem. It starts off by noting that there is no legal mandate for performance testing, moves to its role in program evaluation and jumps to an acknowledgment of a relationship between competency education and performance testing. From there, the subsection moves to a brief discussion of student certification and the legal implications. All of these are areas in which issues exist that should be identified and discussed. However, the issues do not fit well beneath the heading "Legal Mandates," and they lack supportive context due to this organization.

The subsection titled "Human Resources" has three paragraphs on that topic, but a final two-sentence paragraph touches on two key issues in performance testing: (1) the directness of the relationship of the performance measures to the job situation and (2) the need for follow-up (validation) studies. The implications of the crucial first point for cost, validity, legal defensibility, and student utility need to be discussed in detail as does the second point on validity studies.

The legal issues surrounding performance testing are discussed briefly and somewhat inappropriately in the subsection titled "Student Needs." These issues should have been developed at greater length. For instance, the authors could have shown how performance tests tend to require greater job relatedness and validity in testing, but the frequent use of raters requires careful safeguards so that bias does not creep in.

The subsection on "Institutional and Curriculum Concerns" touches on the key issues of cost and time required for performance testing, but it fails to offer help in deciding when the greater cost and time is justified.

In the section titled "Responses of Vocational Education to the Issues," they note that performance testing is not a panacea and there are times when other forms of testing are preferable. The section would have been more comprehensive if they had also said something about performance testing being only one more instrument in the growing arsenal of instruments for pupil and program evaluation, and about its place in a balanced and comprehensive evaluation strategy.

Spillman and Wade seem to support the notion that the chief contribution of performance testing in vocational education will be to program accountability rather than pupil assessment. I believe that it is a mistaken notion since the needs of accountability are already well served by simpler and less expensive measures such as the proportion of graduates who obtain and retain jobs, rating of job supervisors, and so forth. I believe that it is in the areas of student needs and job preference identification, instructional planning, and certification that performance testing will make its major contribution.

The other sections in this paper touch on the key issues, including validity and reliability, relative difficulty of development, effects on students, legal concerns, the need for clear task analysis, and the desirability of avoiding mandates of the use of performance tests. Unfortunately, the allusions to these issues are brief.

In summary, both papers attempt to tie together four complex issues underlying performance testing and their implications for vocational education. This is not an easy assignment and the contributors are to be commended for their efforts. This comments paper attempted to highlight and support points made by the contributors and, in some cases, to raise additional points. Taken together, however, the three papers only touch the tip of the iceberg. Vocational education is bound to review these issues time and time again as it designs and implements performance tests. The vocational education system is complex and dynamic, and as it changes, so must its evaluation methods.

CHAPTER SEVEN

GLOSSARY

The handbook is replete with terms that may be unfamiliar to some of the audience. In response, a glossary of important terms appearing in the papers was prepared. The definitions contained in the glossary were drawn from the papers wherever possible. It should be noted that in some cases the same term was defined by more than one author. In those cases, the briefest definition was selected for inclusion.

GLOSSARY OF TERMS

Term	Definition
Behavioral Process	The way in which a task, duty or operation is carried out.
Behavioral Product	The outcome resulting from some form of behavior
Change	Any alteration in the status quo.
Change Advocate	Some person, group or organization acting as an initiator in a "change process."
Common Law	The law of a country or state based custom, usage, and the decisions and opinions of law courts.
Competency Based Education	The usage of competencies (skills, attitudes, values, or appreciation that is deemed critical to successful employment) as a basis for development of curriculum content: this encompasses making available explicit criteria for each competency, assessing competence in applied settings, having demonstrated competence serve as a determiner of student progress, and focusing on facilitation of student achievement of competencies.
Concurrent Validity	The relationship of a test with meaningful samples of behavior as criteria.
Consistency Validity	The extent to which a person's result on a test corresponds to the person's performance on a task which the test presumably assesses when both performances are measured at approximately the same time.
Construct Validity	The extent to which a test measures hypothetical concepts or qualities.
Content Validity	The extent to which the content of the test samples subject matter, skills or behavior which the test attempts to assess or predict.
Contrast Error	The tendency on the part of raters to see others as opposite to themselves.
Criterion-Referenced Tests	Tests in which an individual is assessed relative to a specified standard rather than to his/her performance relative to other individuals or to group norms.
Criterion Validity	The ability of a test to predict future school or job performance.

GLOSSARY

Critical Competencies	Skills identified as essential to adequately perform a specific occupation.
Direct Assessment	Direct observation in a real life setting.
Due Process	An individual's right to be treated with fairness, consistency, and lack of arbitrariness by governmental agencies and employers
Educational Change	Any significant alteration in the status which is intended to benefit the people involved.
Equal Protection	A constitutional principle related to due process, prohibiting any state from denying to any person within its jurisdiction the equal protection of the laws.
Error Variance	The variability of measures due to random fluctuations, having a basic characteristic of self-compensation.
Face Validity	The apparent ability of a test to measure what it appears to measure.
Generosity Error	The error that results when raters overestimate the positive qualities of individuals they like.
Halo Effect	The effect that results when raters generalize their impressions from one rating to another.
Ideas in Good Currency	Ideas which become important by having an impact on the formulation of public policy.
Innovation	A product or practice new to the adopting unit (e.g., school system, classroom).
Mapping Backwards	A technique for arriving at an estimate of what will be needed to successfully implement a new program or practice.
Minimum Competency Testing	A standardized examination designed to demonstrate whether a student has reached a given level of proficiency in any one of several basic academic skills required to function in everyday adult life.
Norm	A standard of achievement as represented by the average achievement of a large group.
Norm-Referenced Tests	A task which seeks to compare an individual's performance relative to the average performance of a group of similar individuals.

Performance Test	1. Refers to tests in which the test stimulus, the desired response, and the surrounding conditions approximate the reality of an actual situation drawn from a specific occupational or role-based content. 2. They assess a portion or all of an actual work setting by attempting to approximate the reality of the actual work setting.
Predictive Validity	The ability of test scores to relate to criterion measures which are based on some future performance.
Prima Facie	At first view, on the first appearance.
Procedural Due Process	The process that requires that the state act in a fair manner when it deprives a citizen of liberty or property.
Reliability	Whether the test measures a characteristic accurately and consistently.
Response Characteristics of Tests	Two response categories have been of defined: I) respondent behavior requires the examinee to choose from a limited set of clearly defined response alternatives; 2) operant responses are characteristic of behavior in real life situations, and hence do not have artificial, preconceived constraints limiting the behavior that might be observed.
Simulation	The process of abstracting some aspects of reality and concretely representing it in the form of a specific simulated task which examinees are expected to perform.
Standard Error of Measurement	An indication of the magnitude of error between "true" and observed performance. The larger the standard error, the less confidence can be placed in the findings.
Standardization	The administration of a performance test to which each student in an identical manner by means of: the provision of a handbook providing directions to both examiner and student; a set of jobs required by each candidate, including information of specific criteria, item insights and the amount of time usually required to complete each subunit of the test; and, a scale stipulating specific criteria.
Street Level Bureaucrats	A government official such as teachers, police officers, welfare workers or public health officers, who interact directly with the public and make decisions on the basis of individual initiative as well as established routine.
Stimulus Characteristics of Tests	A test which contains a set of instructors, a prompt, a demand, or an event that initiates the examinee's behavior.

GLOSSARY

Substantive Due Process	The process that requires that the action of the state be rational and reasonably related to a legitimate state objective
Surrounding Conditions	The environmental conditions under which a task is performed.
Targeted Consumer	Those consumers to whom the educational innovation is directed
Test Bias	The characteristic of a test in being free of various types of content bias (e.g., numerical, role, status, stereotype and familiarity.)
Validity	Refers to whether the test actually measures the characteristic that it claims to measure.
Verisimilitude	Performance tests in vocational education which take the form of work samples or job simulations closely resembling the actual on-the-job task to be performed by the worker; those tests having a high face validity.
Work Samples	Selected tasks performed under controlled environmental and time conditions. The aim is to standardize tasks and enhance replicability across examinees under conditions controlled and specified by the examiner.

CHAPTER EIGHT

CONTRIBUTORS

The contributors to this handbook were drawn from varied disciplines and professions in an effort to address the issue of performance testing from a multidiciplinary perspective. Thus, while the names and professional affiliations of some of the contributors may be familiar, others may not. To provide a context for the reader, Chapter Eight consists of a brief biographical sketch of each contributor.

CONTRIBUTORS

Henry Borow (Ph.D., Pennsylvania State University) is a professor of psychological studies, General College and College of Education at the University of Minnesota. He is the author of over 100 journal articles, books, book chapters, and tests. Dr. Borow is a past-president of the National Vocational Guidance Association and editor of its fiftieth anniversary volume, *Man in a World at Work* (1964). He was a postdoctoral fellow of the American College Testing Program and served on the national advisory board of the National Center for Resarch in Vocational Education.

William G. Buss (L.L.B., Harvard University) is a professor at the University of Iowa College of Law. He has published extensively in the areas of educational law and constitutional law.

Curtis R. Finch (Ph.D., Pennsylvania State University) is professor and chairman, General Vocational and Technical Education, Virginia Polytechnic Institute and State University. He has served on the faculties of Ohio State University and Pennsylvania State University. Dr. Finch has served as editor of the *Journal of Vocational Education Research* and *Occupational Education Forum. He has authored or co-authored over seventy professional articles, papers, and reports and is co-author of Curriculum Development in Vocational and Technical Education* (Allyn and Bacon, 1979). Dr. Finch served as a Senior Fulbright Lecturer to Cyprus during the first part of 1980.

Raymond S. Klein (Ed.D., State University of New York, Buffalo) is the program coordinator at the National Occupational Competency Testing Institute (NOCTI), Albany, New York. He has also served on the faculty of Pennsylvania State University and as the director of research for the New York State Department of Education.

Samuel A. Livingston (Ph.D., Johns Hopkins University) is a program research scientist at the Center for Occupational and Professional Assessment at the Educational Testing Service. He has been involved in the area of performance testing for the past seven years during which he has developed performance tests for such varied occupations as firefighters, radiologic technicians, dental assistants, dental hygienists, and machine tenderers.

H. Brinton Milward (Ph.D., Ohio State University) is an assistant professor of Business & Public Administration at the University of Kentucky. He formerly served as associate director of the Graduate Program in Public Administration at the University of Kansas. His published research has been in the fields of organization theory and public policy. Dr. Milward is currently testing an organizational theory of discrimination in colleges and universities. He also serves on the editorial board of *The Annals of Public Administration.*

Evelyn Perloff (Ph.D., Ohio State University) is an associate professor of Nursing Research and of Psychology in the School of Nursing at the University of Pittsburgh. She has also served as a faculty member at Purdue University, Northwestern University, and Kendall College. Dr. Perloff has published widely in the area of program evaluation.

CONTRIBUTORS

Diana C. Pullin (J.D., Ph.D., University of Iowa) is a staff attorney at the Center for Law and Education Inc., Cambridge, Massachusetts. She has previously served as legal counsel for local school districts and an intermediate educational agency. Dr. Pullin represented the students and parents who successfully challenged the State of Florida's use of a minimum competency test to deny high school diplomas in the federal court lawsuit *Debra P. v. Turlington*. Dr. Pullin's previous publications have been in the areas of minimum competency testing and the law relating to the education of children with special education needs.

Marvin R. Rasmussen (M. Ed., University of Oregon) is Director of District Programs for the Portland (Oregon) public schools. He has served as the director of career education programs for the Portland Public schools and as a principal, administrative vice principal, and secondary teacher.

Stephen J. Slater (Ph.D., University of California at Santa Barbara) has been responsible for coordinating activities of the Clearinghouse for Applied Performance Testing (CAPT) an NIE sponsored project at the Northwest Regional Educational Laboratory. In that position, Dr. Slater edited the *CAPT Newsletter*, prepared an extensive annotated bibliography on applied performance testing, and organized the 1979 Annual CAPT Conference, entitled Alternative Conceptions of Competence Assessment. Recently, Dr. Slater joined the staff of the Planning and Evaluation Section, Oregon Department of Education.

Robert E. Spillman (M.S., University of Kentucky) is Director of the Kentucky Bureau of Vocational Education. He has served in the capacities of acting deputy superintendent for Occupational Education, secretary to the State Board for Occupational Education; and director of Supporting Services Division in the Bureau of Vocational Education. Mr. Spillman has been a secondary vocational teacher, teacher educator, and curriculum writer. He has articles on competency-based vocational education. In addition, he has been Kentucky's representative on the V-TECS Board of Directors serving as chairman of the organizing committee and Board chairman for three years. Mr. Spillman and Dr. Wade have jointly been involved in several other related activities. They were co-directors of a 1975-76 Region IV EPDA Workshop on CBVE and co-authors of articles of CBVE and vocational student organizations. They participated in the study, design, and implementation of one of the most comprehensive statewide programs of CBVE. Kentucky's program, based on the V-TECS catalogs, currently involves 22 occupational areas and has been implemented in 1,090 specific programs.

Janet E. Spirer (Ph.D., Ohio State University) is a research specialist at the National Center for Research in Vocational Education. Dr. Spirer served as director of the project under which the handbook was produced and edited the manuscript. Her research interests focus on human resource policy and program evaluation.

John F. Thompson (Ph.D., Michigan State University) is a professor and chairman of the Department of Continuing and Vocational Education at the University of Wisconsin-Madison. His research interests and publications have been in the areas of philosophy of vocational education, curriculum in vocational education and inservice and professional development education.

Nellie Carr Thorogood (M. Bus. Ed., North Texas State University) is Director of Occupational Education and Technology at San Antonio College. She has community college and university work experiences as an instructor, cooperative education coordinator, program area coordinator, division chairperson, and teacher-educator. She has served as a member of the Alamo Consortium Private Industry Council and Youth Council; as a chairperson of a statewide committee studying meeting the special needs of occupational students in Texas; as a member

of both state and national task forces on the impact of vocational education data systems on postsecondary occupational education, and as an advisory member of several local employment and education programs.

Paul L. Tractenberg (J.D., University of Michigan) is a professor of law at Rutgers School of Law in Newark, New Jersey where he specializes in public education law. Within that field, he has taught courses and seminars at the law school, researched and written extensively, and presented many papers and speeches to national, regional and statewide organizations. Prof. Tractenburg has also consulted with many groups and established an ongoing public interest law center to represent the interests of students and parents. Currently, he is especially involved in assessing the legal implications of minimum competency and performance testing of students, teacher competency measures, and school finance reform. Also, he is writing a book, under a Ford Foundation grant, about the role of the courts in educational reform.

Charles D. Wade (Ed.D., University of Kentucky) is the director of the Division of Vocational Program Development of Education (Kentucky Bureau of Vocational Education). He has served as an RCU research associate, a program supervisor, a secondary vocational teacher, and a part-time teacher educator. Dr. Wade has addressed a variety of national and state conferences on such topics as program planning, competency-based curriculum, cooperative education, and evaluation of vocational programs.

Jack C. Willers (Ph.D., University of Texas at Austin) is a professor of history and philosophy of education at George Peabody College for Teachers of Vanderbilt University. He has held a Fulbright-Hays Lectureship Award to Iran, Greece and Egypt. Dr. Willers has published widely in several journals on philosophy and the social foundations of education and educational policy issues.

EVALUATION PUBLICATIONS OF THE NATIONAL CENTER FOR RESEARCH IN VOCATIONAL EDUCATION ON EVALUATION

EVALUATION HANDBOOKS SERIES

- Guidelines and Practices for Follow-up Studies of Former Vocational Education Students
- Guidelines and Practices for Follow-up Studies of Special Populations
- The Case Study Method: Guidelines, Practices, and Applications for Vocational Education
- Performance Testing: Issues Facing Vocational Education
- Evaluation Guidelines and Practices for State Advisory Councils

CAREER EDUCATION MEASUREMENT SERIES

- Assessing Experiential Learning in Career Education
- Career Education: A Compendium of Evaluation Instruments
- Improving the Accountability of Career Education Programs: Evaluation Guidelines and Checklists
- A Guide for Improving Locally Developed Career Education Measures
- Using Systematic Observation Techniques in Evaluating Career Education

VOCATIONAL EDUCATION OUTCOMES SERIES

- Viewpoints on Interpreting Outcome Measures in Vocational Education
- Vocational Education Measures: Instruments to Survey Former Students and Their Employers
- Vocational Education Outcomes: An Evaluative Bibliography for Empirical Studies
- Vocational Education Outcomes: Perspective for Evaluation
- Vocational Education Outcomes: A Thesaurus of Outcome Questions
- Vocational Education Outcomes: Annotated Bibliography of Related Literature

For Information concerning the above publications, please contact:

Program Information Office
The National Center for Research
in Vocational Education The Ohio State University
1960 Kenny Road
Columbus, Ohio 43210
(614)486-3655

"END OF DOCUMENT"